

ABSTRACT

Title of Dissertation: HOW DO RESOURCE AVAILABILITY,
INFORMATION FLOW AND TARGETED
INTERACTIONS SHAPE ECO-
EVOLUTIONARY PROCESSES IN
MICROBIAL ORGANISMS ACROSS
SCALE?

Anshuman Swain, Doctor of Philosophy, 2022

Dissertation directed by: Prof. William F. Fagan, Department of Biology

This work explores the crucial link between biological organization and dynamics in microbes. I have developed and applied different sets of quantitative tools to investigate this theme using a multi-disciplinary approach.

My first project uses flow optimization tools to explore microbial intra-cellular dynamics as a function of resources, cell structure, and trade-offs between major energy production processes. This model explains emergent **cell-level** properties such as the dependence of bacterial growth rates on cell shape, size patterns in long-term evolutionary experiments, and the intriguing Warburg effect – wherein cells use inefficient anaerobic energy production pathways instead of efficient aerobic processes, despite ample oxygen availability.

Moving to the **community-level**, I explored the mechanisms by which high levels of diversity are maintained in natural systems. Via a cutting-edge agent-based model that incorporates biologically realistic higher-order interactions, a continuous species space in which microbes interact via antibiotics and defenses, and an evolution-via-mutation regime, I link the trajectory of community assembly to both spatial heterogeneity patterns and diversity dynamics, thereby providing a new perspective on the origin and maintenance of diversity in complex communities.

At the **ecosystem-level**, I built a dynamical system model to explore how the availability of resources and oxygen—mediated by interactions among microbial functional groups—affect environmental production of major greenhouse gases, like methane, especially on early Earth. This work provides 1) the first ecosystem-based functional explanation of periodic global methane ‘hazes’ in Earth’s early atmosphere and 2) a way to predict oxygen concentrations of ancient environments in a novel way.

Lastly, in research that can involve multiple scales of biological organization, I use novel network science and information theoretic metrics to understand differences in interaction patterns in networks through the lens of degeneracy, noise, and determinism. The tradeoff between degeneracy and determinism can help us define the most informative scale of a network, and this is explored in the project through the development of new openly accessible toolkits.

HOW DO RESOURCE AVAILABILITY, INFORMATION FLOW AND TARGETED
INTERACTIONS SHAPE ECO-EVOLUTIONARY PROCESSES IN MICROBIAL
ORGANISMS ACROSS SCALE?

by

Anshuman Swain

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:

Professor William F. Fagan, Chair
Professor Alan J. Kaufman
Professor Karen Carleton
Professor Margaret A. Palmer
Professor Michelle Girvan

© Copyright by
Anshuman Swain
2022

Foreword

With the approval of the dissertation committee chair, Dr. William F. Fagan and other members of the dissertation committee, Chapters 2, 4 and 5 of this dissertation are included as previously published research. The citations for the publications are provided at the beginning of those chapters and are also mentioned here:

Chapter 2: Swain, A. and F Fagan, W., 2019. A mathematical model of the Warburg Effect: Effects of cell size, shape and substrate availability on growth and metabolism in bacteria. *Mathematical Biosciences and Engineering*, 16(1).

Chapter 4: Swain*, A., Fussell*, L. and Fagan, W.F., 2022. Higher-order effects, continuous species interactions, and trait evolution shape microbial spatial dynamics. *Proceedings of the National Academy of Sciences*, 119(1).

(*equal contributions, first co-authors)

Chapter 5: Klein*, B., Swain*, A., Byrum, T., Scarpino, S.V. and Fagan, W.F., 2022. Exploring noise, degeneracy, and determinism in biological networks with the einet package. *Methods in Ecology and Evolution*. <https://doi.org/10.1111/2041-210X.13805>

(*equal contributions; first co-authors)

Also, Chapter 3 is available as a preprint on BioRxiv as:

Swain, A., Kaufman, A.J., Kalinowski, M., Yarwood, S.A. and Fagan, W.F., 2021. Were Neoproterozoic atmospheric methane hazes and early Paleoproterozoic glaciations driven by the rise of oxygen in surface environments?. *bioRxiv*, pp.2020-10.

Dedication

For my Māmā and Bābā,
who have supported and tolerated me through all my ups and downs;

and

For my mentors throughout my life,
who stood by me, even when I was making difficult transitions, and was not thinking straight.

Acknowledgements

I am grateful to my advisor Dr. Bill Fagan for supporting me throughout my doctoral journey and helping me see and explore the world in a new light. The level of independence that Bill gave me, enabled me to find what fields of research I like or dislike and investigate new questions and frameworks that might not have been possible through the standard project-focused research paths. I am fortunate to be a part of the COMBINE (COMputation and Mathematics for BIOlogical NEtworks) program at the University of Maryland (UMD), which transformed my research focus and helped me in finding other people at UMD who were interested in complex systems and allowed me to discover the larger global complex systems community. I would like to thank the team behind it – Dr. Michelle Girvan (a member of my dissertation committee), Dr. Daniel Serrano, and Glynis Smith – for fostering such a welcome and supportive program space. I must mention that through COMBINE, I got introduced to the Santa Fe Institute (SFI) Complex Systems Summer School (CSSS) in 2019, which expanded my research worldview.

Working in multiple collaborations among early career scientists from across a wide breadth of disciplines at SFI CSSS, I discovered important aspects of project management/planning and proper communication across disciplinary boundaries. My experience at SFI CSSS led me to two more international complex systems workshops – CNWW (Complex Networks Winter Workshop) 2019 and WWCS (Winter Workshop in Complex Systems) 2020, and I am thankful to all the three. Most importantly, two of my dissertation chapters (Ch. 4 and 5) were a result of friendships and collaborations that I fostered through these workshops.

Other members of my committee played a critical role in developing my knowledge base, and navigating my doctoral life (in alphabetical order of surnames): Dr. Karen Carleton, who would always hear me out and support me with resources whenever I needed any advice or help;

Dr. Alan Jay Kaufman, who introduced me to the wonderful world of geochemistry and paleo-ecosystems; and Dr. Margaret Palmer, who welcomed me into her lab meetings and helped me understand microbial interactions better.

During my initial time at UMD, Dr. Philip Johnson and Dr. Philip Staniczenko were always there for me whenever I had any questions about quantitative methods or if I needed any quick professional advice. In near future, I plan to switch to working on paleobiology (through the eyes of complex systems), and this decision was a result of many projects that stemmed from my remarkable interactions with Dr. Conrad Labandeira, Dr. Ellen Currano, and Dr. Nick Butterfield.

The unwavering support from my friends helped me sail through the graduate school years and made me a better person. The list is long, but I would like to specifically thank (in alphabetical order of surnames): Zabdiel Alvarado Martinez, Lauren Azevedo Schmidt, Nicki Barbour, Kevin Bennett, Domenick Braccia, Mary Chey, Levi Fussell, Xiaozhen Fu, Prakhar Gupta, Dr. Brennan Klein, Dr. Gussie Maccracken, Dr. Julie Mallon, Naren Manjunath, Dr. Jesse Milzman, Nikhil Oberoi, Spandan Pathak, Nick Rachmaninoff, Vanessa Rubio Ramos, Subhayan Sahu, Hadi Vafaei and Dr. Jake Weismann.

Lots of love to the whole Behavior, Ecology, Evolution and Systematics (BEES) community at UMD, who have made the last 5 years incredibly memorable for me. It is the most fun and understanding community that I have ever been a part of.

I am grateful to Srishti Sood, who has always stood by my side, (endured hours of me babbling about my research, random ideas and interesting facts) and has patiently taught me many critical things that have improved my work and professional life.

Lastly, I would like to mention my family, who have made me who I am today and without whom I could not have achieved much. Māmā (Jayashree Swain), Bābā (Dr. Ranjan Kumar

Swain), and Ajā (Benudhar Samantaray) have always encouraged me to pursue my dreams, sacrificed a lot for my sake, have been ever-present whenever I needed shoulders to stand on and did everything they could, to help me achieve it. Ā'i (Sobhabati Samantaray), Jējēmā (Malatilata Swain), Jējē (late Bhagyadhar Swain), and Māun'i (Manjushree Ray) have always been a constant source of support and goodwill. Thank you for always being there for me!

Note: My graduate research and travel were funded by:-

The Ann G. Wylie Dissertation fellowship

The University of Maryland CMNS Dean's Fellowship

Devra G. Kleiman Memorial Graduate Scholarship

National Science Foundation - COmputation and Mathematics for Biological Networks (NSF-COMBINE) Fellowship

Table of Contents

Foreword	ii
Dedication	iv
Acknowledgements	v
Table of Contents	viii
List of Tables	xi
List of Figures	xii
Chapter 1 : Introduction	1
Complex Systems Thinking in Biological Systems	1
Microbes as complex systems	2
Short Overview of the Tools and the Questions	4
Linear optimization for understanding metabolic reactions (intra-cellular level)	4
ODE models for describing dynamics between microbial guilds (ecosystem level)	4
Individual-based models to explore microbial community assembly (community level)	5
Network techniques to look at higher order information in interactions (multi-scale)	7
Conclusion	8
References	9
Chapter 2 : A mathematical model of the Warburg Effect: Effects of cell size, shape and substrate availability on growth and metabolism in bacteria	15
Abstract	15
Introduction	16
The Model	21
Results	28
Dependence on substrate concentration	28
Dependence on cellular size and shape	29
Overflow metabolism and bacterial metabolic regime	30
Discussion	40
Comparison with previous work	40
Support from previous studies and experiments	43
Further comments	45
Acknowledgements	45
References	46

Chapter 3 : Were Neoproterozoic atmospheric methane hazes and early Paleoproterozoic glaciations driven by the rise of oxygen in surface environments?	52
Abstract	52
Introduction	52
Results	59
No External Forcing	59
External Forcing	62
Discussion	66
Conclusion	73
Modeling Microbial System Dynamics	73
Notation for Sulfur Isotope Compositions	78
Acknowledgements	79
Code and data availability	79
References	80
Supplementary materials	85
Chapter 4 : Higher-order effects, continuous species interactions, and trait evolution shape microbial spatial dynamics	109
Abstract	109
Introduction	110
Results	116
Discussion	124
Methods	130
Model Description	130
Model Assumptions	133
Model Implementation	136
Data Analysis	138
Acknowledgements	139
Code and Data Availability	139
References	140
Supplementary materials	150
Chapter 5 : Exploring noise, degeneracy, and determinism in biological networks with the ‘einet’ package	157
Abstract	157
Introduction	157
Overview of the metrics	160

Breaking down effective information: Determinism and degeneracy	162
Causal emergence: Informative higher scales	164
Implementation and usage	165
Example usage (R)	167
Effective information and effectiveness	167
Causal emergence	167
Example usage (Python)	168
Effective information and effectiveness	168
Causal emergence	168
Software and Code Availability	169
Acknowledgements	169
References	171

List of Tables

(Main tables by chapter)

Table 2.1: Trade-offs involved in Warburg effect	21
Table 3.1: Differential equation parameters governing the dynamics of a system of five microbial functional groups. Parameters summarize the efficiency of resource incorporation, the efficiency of resource utilization, and the energy requirements for homeostasis.	75
Table 4.1: Brief description of model parameters.....	131

List of Figures

(Only main text figures listed, by chapter)

Figure 2.1: A simplified representation to model Warburg effect. The external glucose is imported into the bacterial cell via glucose transporters located on the cell membrane at a rate r_0 . This glucose is utilized by upper glycolysis pathways to convert it to pyruvate at a rate r_1 . This converted substrate is then divided into two paths: either to OP (at a rate r_2 at the cytochrome oxidase located in the membrane) or to fermentation/aerobic glycolysis (AG) (at a rate r_3 at the fermentation enzyme sites located in the cytoplasm). Note that only the substrate used in the latter shunt goes into biomass production whereas substrate used in OP is lost as CO_2 23

Figure 2.2. Representation of the solutions for a bacterial cell at steady state with all features except glucose concentration constant. The dot-dashed red line represents relation (9), the solid green line represents relation (10) and the dashed blue line represents Equation (11). The area shaded in yellow represents the solution for (9) and (10), and the red dot represents the optimized value of r_3 for (9)-(11). 31

Figure 2.3. Representation of the solutions for a bacterial cell at steady state with all features constant except the ratio of surface area to volume of the cell (denoted by Z). The dot-dashed red line represents relation (9), the solid green line represents relation (10) and the dashed blue line represents Equation (11). The area shaded in yellow represents the solution for (9) and (10), and the red dot represents the optimized value of r_3 for (9)-(11). 32

Figure 2.4. Representation of overflow metabolism in a bacterial cell. (a) The dot-dashed red line represents relation (9), the solid green line represents relation (10) and the dashed blue line

represents Equation (11) and the solid purple line represents relation (5). The area shaded in yellow represents the solution for (9) and (10), and the red dot represents the optimized value of r_3 for (9)-(11). The black dot represents altered final solution due to (5). See text for details. (b) The solid green line represents (10), the solid purple line represents the final solutions of (5), (9)-(11) and the black dot dash line represents (12). Δr_3 represents the estimate of overflow metabolic products. The phase space has been divided in four parts based on the solid purple line and (12). (c) Overflow products vs. growth (Equation 7)..... 34

Figure 2.5. A representation of the effect of overflow metabolism in temperate zone wetlands. **A** refers to heterotrophic bacteria, **B** refers to methanogens (which feed on lactate and acetate to produce methane), and **C** refers to methanotrophs (which consume CH_4 and O_2 to produce CO_2). The area shaded yellow in the figure, refers to the CO_2 produced by methanotrophs, by oxidizing CH_4 . The CO_2 and methane emissions are representative from previous literature (Corteselli2017). 39

Figure 3.1. Flow diagram representation of the model system. Here (in blue) A, B, C, F, G and H are functional groups of heterotrophic bacteria, methanogens, aerobic methanotrophs, fermenters, anaerobic methanotrophs, and anaerobic heterotrophs, respectively. The flux arrows indicate processes that produce (in red) either metabolic reactants or gases cycled among the microbial functional groups. Ordinary differential equations with detailed explanations representing this dynamical system appear in **Methods**..... 57

Figure 3.2 Dynamic variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with the coefficient of CH_4 production (β) set at 1, 10, and 50% in rows 1-3, respectively. Note that the resource (R) axis begins at 2; this is due to

constraints posed by the minimal energy requirements for each microbial functional group,
below which all functional groups diminish. 61

Figure 3.3: Dynamic variations in O₂ and CH₄ for the externally forced model with constant rates of addition of R and O₂. The O₂ input rate (σ_2) varied between 1% to 50% relative to the resource input rate across panels (σ_1), but in all cases the R input rate (σ_1) was held constant at 1. 64

Figure 3.4: (A) shows the dependence of amplitude (in red) and time period (in black) of the CH₄ oscillations in Fig. 3.3 as a function of the O₂ input rate (σ_2). (B) demarcates the resource input (σ_1) -oxygen input (σ_2) space on the basis of oscillatory behavior for our standard set of parameters. (C) depicts the dependence of the slope of the line separating the space between oscillatory and non-oscillatory behavior in σ_1 - σ_2 space on specific parameters (for description of parameters, see **Methods**; **Table 3.1** in **Methods**) over a wide range of parameter combinations through partial rank correlation coefficient (PRCC). 65

Figure 4.1: A conceptual representation of the model, with descriptions of all the parameters involved. See Methods for details about implementation. 115

Figure 4.2: System dynamics in three example conditions of initial conditions pertaining to different regimes of spatial disturbance behavior (Low, Medium and High) (rows A-C). Snapshots of each simulation at different timesteps (10, 100, 250 and 500) are shown in columns 1 through 4. The respective species dynamics (where different colors represent different species ‘bins’, see Methods) and diversity dynamics are represented in columns 5 and 6, respectively. In columns 5 and 6, the time taken for the simulation to stabilize (i.e., the community formation

time or CFT) is denoted with a vertical red dotted line. The community existing after the system has reached the CFT is termed an eco-evolutionary stable community (ESC). 117

Figure 4.3: Dependence of model dynamics on different parameters and community formation time (CFT). (A) shows the variable importance score for all the parameters from a random forest classification (with 10,000 trees) on the three categories of spatial disturbance (OOB=10.2%). (B) and (C) denote the variable importance scores from random forest regression models (with 10,000 trees) for Shannon diversity (SD) mean and variance respectively. (D)-(E) show the SD mean-variance plots overlaid with mutation size, growth radius, and community formation time (CFT) values of the respective simulations. For (A)-(C), IncNodePurity (used in regression random forests) and MeanDecreaseGini (used in classification random forests) refer to how well a parameter explains the prediction variable in question. 118

Figure 4.4: Effects of Community formation time (CFT) on the spatial diversity dynamics. (A) depicts the four optimal clusters (groups) of SD mean-variance space found using k-means clustering (see Figure S3). (B) shows the results from the partial rank correlation coefficient (PRCC) analyses of the four groups and the overall data. (C) shows the proportion of spatial disturbance regimes present in different groups. (D) and (F) show the distribution of the values of mean and variance of Shannon Equitability and Simpson's index, respectively, for a smaller subsample of the runs, with points colored by groups from (A). (E) depicts the group-wise histogram of Shannon equitability and (G) the histogram of Simpson's index (for clarity, only the range between 0 and 0.04 is shown). 121

Figure 5.1: Determinism and degeneracy. Simple networks highlighting the differences between high/low determinism and degeneracy. (a) In this network, random walkers on node A

will traverse to node B with probability of 1.0. **(b)** In this network, random walkers on node A will traverse to any other node in the network with probability $1/n$, where n is the size of the network. **(c)** In this network, no single node has a higher in-weight than any other node, which means that degeneracy has been minimized. **(d)** In this network, every node has a single outgoing edge, each of which connects to node A; this is an example of a maximally degenerate network 161

Figure 5.2: Causal emergence in protein networks. To illustrate the basic steps for calculating causal emergence in networks, we use the protein-protein interaction network of *Mycoplasma putrefaciens* as an example. **(a)** The original (micro scale) network. **(b)** The different color circles are meant to highlight subgraphs in the original network that are especially noisy—either because they consist of densely connected nodes or because they are sources of degeneracy. **(c)** In order to calculate causal emergence, we construct a macro scale network by grouping the subgraphs from (b) into macro nodes, while preserving the total in/out edge weights into the selected subgraph. Note here that the EI of the network in (a) and the EI from (c) are different: the macro scale network has more effective information than the original micro scale network. 166

Chapter 1 : Introduction

Complex Systems Thinking in Biological Systems

Organization in a biological system is shaped both by constraints from factors outside the system, and by interactions among elements within it. Unlike physical systems where the interactions between its components are usually ‘homogeneous and isotropic’ (the nature of interaction does not depend upon component identity, e.g., gravitational forces between any two objects can be expressed in the same way), interactions between components of a biological system (and in many cases, constraints affecting them) are specific in nature, and can evolve through time (Turner et al., 2018). In addition, the components in a system themselves can evolve due to the interactions and constraints placed upon them. Therefore, in biological complex systems, the interactions between components can sometimes result in global behaviors that are not predictable from the known properties of the components alone (von Bronk et al., 2018).

For many biological systems and the phenomena associated with them, a major challenge is to discern an appropriate scale of description and whether the description is a sufficient representation of the system’s complexity. This is further complicated by the fact that all biological systems are spread across a hierarchy of description levels that researchers have constructed to categorize, understand, and explore natural systems in an easier way (Morrison, 2015).

Modeling systems consisting of many interacting components, spread across multiple scales, involves making assumptions about the system through our model choices. Most models serve the reductionist paradigm, although to different degrees, for ease and convenience of representing complex systems (Regenmortel, 2004). Boiling down a system to its most crucial components is the major backbone in modeling a phenomenon, but the degree of reductionism

plays an important part in the validity of the model. Complex systems thinking entails keeping the heterogeneous, non-isotropic nature of interaction structures and the relationship between scale as intact as possible to generate honest representations of real (biological) systems (Turner et al., 2018).

Such ‘complex system’ models help us to identify properties of complex systems that are otherwise lost in simpler representations (Ladyman et al., 2013). These properties can include, but are not limited to, emergence (complex systems as a whole can be very different, and often unexpectedly, from what is expected of their individual components), unpredictable long-term behavior (such as cyclic behavior or chaos), self-organization or structuring (systems can sometimes generate non-trivial patterns spontaneously in absence of a blueprint), adaptation and evolution (of components) (Auyang, 1998; Ladyman, 2013).

Microbes as complex systems

Stating that microbes are everywhere around us, might be an understatement. Microbes are found in and on our human bodies, and that of all animals and plants that we observe, in the atmosphere, deep in the ocean, and under the earth’s crust. They were the earliest inhabitants of our planet (Golding and Glikson, 2010) and continue to dominate the diversity of forms around us, evolving and adapting to the surroundings every moment. They occupy two domains of life (bacteria and archaea) and are the cause of our (eukaryotic) very existence (López-García and Moreira, 2020). The estimate of microbial diversity runs into a figure of a trillion species (Locey and Lennon, 2016), and even a gram of soil may contain hundreds of millions of microbial cells and thousands of species. These microbial cells do not just inanimately exist – they interact intricately with the environment and with other microbes. They are metabolic marvels, and can synthesize, process, and modify an incomprehensible array of chemicals, based on the availability of substrates in their

surrounding medium and external environmental factors (Fuhrman, 2021). The effects of these metabolic reactions are both local and global at the same time: these affect the cellular level conditions of each microbe and also, as an aggregate, impact the planetary scale biogeochemical cycles of the Earth (Rousk and Bengtson, 2014). These metabolic activities also result in chemicals that aid microbes as they interact with other microbes in a variety of ways – they cooperate with each other for sustenance, compete for resources, and kill for space – and most of these interactions are specific in nature (e.g., between a fixed set of microbial partners) and happen under only a limited set of circumstances (e.g., environmental conditions such as resource scarcity) (Peterson et al., 2020; Seth and Taga, 2014; West et al., 2006). Even more, these interactions and the microbially produced chemicals that mediate them, evolve fast – as the microbes reproduce quickly (compared to eukaryotic organisms) and can exchange genetic materials with each other using horizontal gene transfer (Smets and Barkay, 2005). This results in an overlapping of not only different scales of space (cellular to ecosystem) where the interaction plays out and evolves, but also scales of time – where ecological and evolutionary processes happen over similar timespans.

These features make microbes an excellent choice for studies using a complex systems perspective. Although the number of questions that can be explored using such a philosophy in the realm of microbial dynamics are endless, in this dissertation, I focus on a few yet diverse aspects of microbial complex systems. In specific, I look at how interactions at different scales of biological organization affect eco-evolutionary dynamics at various spatiotemporal scales. This interaction between biological organization and microbial dynamics is a crucial problem in biology, and each of my projects looks at a different scale (both spatial and temporal) of microbial dynamics. Each project also uses a different suite of complex systems tools that best represent the scale of the problem.

Short Overview of the Tools and the Questions

Linear optimization for understanding metabolic reactions (intra-cellular level)

Metabolic reactions within a cell at stable state are subject to certain constraints due to their stoichiometry. Therefore, frameworks for studying such reactions typically use constrained linear optimization tools to understand growth and substrate usage in microbes (Orth et al., 2011). Most of these models either include detailed chemical reactions occurring within the cell of a particular species of microbe) or portray a very general perspective of game-theoretic agents in a population with basic metabolic properties (Kareva, 2011; Pfeiffer et al., 2001; Shlomi et al., 2011; Zhuang et al., 2011). Previous works have rarely considered the allometric and (in many cases,) the kinetic constraints on the major reactions affecting metabolism and growth in microbes (Schuster et al., 2015). Keeping these things in mind, my first dissertation project focuses on how resource availability, along with cell size and shape, control growth and metabolism in microbes. Using tools from earlier linear (flow) optimization models, I explore the intra-cellular dynamics as a function of resources, cell structure, and trade-offs between major energy production processes (Swain and Fagan, 2019). My model explains numerous emergent properties of microbes such as cell-shape dependence of growth, size patterns in long term evolutionary experiments and the intriguing Warburg effect – where cells use an inefficient anaerobic energy production pathway instead of efficient aerobic processes, despite ample oxygen availability.

ODE models for describing dynamics between microbial guilds (ecosystem level)

The second dissertation project builds off the first chapter (cell-level properties) to construct an ecosystem-level dynamical system model where different major functional guilds of microbes interact in a targeted way (Swain et al., 2021a). This model uses a suite of ordinary differential

equations (ODEs) to portray the dynamics of the system. This work focuses on how the availability of resources and oxygen, mediated by interactions among microbial guilds, affects environmental production of major greenhouse gases like methane, and causes important geological events like mass extinctions (see Rothman et al., 2014) and methane hazes (see Eigenbrode and Freeman, 2006; Farquhar et al., 2013; Izon et al., 2015; 2017, Williford et al., 2016; Zerkle et al., 2012). I explore an array of different scenarios of resource and oxygen availability through the lens of different microbial processes, which are encoded as parameters in our model. The project then aims specifically on providing the first ecosystem-based functional explanation of 2.5 billion years old global methane hazes (large methane fluctuations) preserved in geochemical records (see Eigenbrode and Freeman, 2006; Farquhar et al., 2013; Izon et al., 2015; 2017, Williford et al., 2016; Zerkle et al., 2012) and allowing researchers a way to predict oxygen concentrations of ancient environments in a novel way.

Individual-based models to explore microbial community assembly (community level)

The third dissertation project takes a more general approach to targeted interaction among microbes, to target a long-standing open question in theoretical biology – modeling the origin and maintenance of species diversity (Chesson, 2001; Hutchinson, 1961). The assembly and maintenance of microbial diversity in natural communities, despite the abundance of toxin-based antagonistic interactions, presents major challenges for biological understanding (Peterson et al., 2020). A usual way for investigating such interactions involves cyclic dominance games with pairwise-mode interactions (Kerr et al., 2002; Kirkup and Riley, 2004; Szolnoki et al., 2014). The inclusion of higher-order interactions in such models permits increased levels of microbial diversity, especially in communities where antibiotic-producing, sensitive, and resistant strains coexist (Kelsic et al., 2015). However, most such models involve a small number of discrete

species, assume a notion of pure cyclic dominance, and focus on low mutation rate regimes (Szolnoki et al., 2014; Kelsic et al., 2015), none of which well represent the highly interlinked, quickly evolving, and continuous nature of microbial phenotypic space (Tarnita, 2018). The incorporation of these features is difficult in traditional models, because of a number of factors. The most important of these is that these models are usually equation-based, and controlling so many parameters can quickly evolve into chaos (see Kelsic et al., 2015). Therefore, instead of using the older equation-based frameworks, we opted for cutting-edge individual-based (or agent-based) models (IBMs). These models are algorithmic or rule-based in nature, and can better account for spatial heterogeneity, individual interactions and stochasticity (Grimm et al., 2005). These models have been used in ecology to account for the complex rule-based characteristics of the various spatio-temporal scales of the ecosystem, ranging from foraging and decision making to landscape-scale changes and conservation (An et al., 2021; Grimm et al., 2005; McLane et al., 2011). In this and previous work (Swain et al., 2021b; Swain et al., 2022), I introduce statistical learning tools to make sense of the parameters, which is usually the missing component when compared with easily explorable equation-based systems. This allows me to identify which parameters most affect the system, and to what extent.

Via IBMs in this project, I incorporate emergent higher-order interactions among microbes, a continuous species space and an evolution-via-mutation regime in this model, through the lens of non-cyclic dominance (i.e., where a single species is not dominant in any specific ecological trait). In this work, I link the trajectory of community assembly to both spatial heterogeneity patterns and diversity dynamics – providing a new perspective on eco-evolutionary dynamics. I find that mutation size and mobility enhanced microbial diversity and temporal persistence to extraordinarily high levels. I also discovered that the duration of the transient phase in community

assembly provides the best guide to equilibrium diversity—highlighting the potentially critical role that antagonistic interactions play in promoting the diversity of bacterial systems.

Network techniques to look at higher order information in interactions (multi-scale)

My fourth dissertation project focuses on a method to look at interactions in a more general sense. Networks are a great way to look at interactions between entities (be it at an organismal, intra-cellular or ecosystem level), with nodes being the entities and edges between them being the interactions (Gosak et al., 2018). Among other uses, these networks can represent host-microbiome interactions, protein-protein interactions within a cell, cellular metabolic pathways, and species interactions, and have revolutionized our understanding of these systems (Faust and Raes, 2012; Gosak et al., 2018; Hoel et al., 2021; Muller et al., 2018). A major problem that arises in many of these networks is trying to understand what scale of the network is the most informative for asking certain biological questions. This is especially compounded by the fact that many biological networks involve inherent noise. Likewise, research tends to focus on the raw, micro-scale network from data/simulations and seldom explores the scale-dependence of properties (Klein & Hoel, 2020).

In this project, we I address this problem by developing novel tools in network science and information theory. I provide a general framework for exploring differences of interaction patterns through the lens of degeneracy, noise, and determinism (see Hoel et al., 2013). The tradeoff between degeneracy and determinism can help us define the most informative scale of a network, and this is the focus of this project (Klein et al., 2022). More interestingly, the process of finding the most informative scale also enables us to find redundant nodes and, in a way, functionally similar node classes. The tools developed in this project are available through software packages

written in R and Python, the two most commonly used programming languages in biological applications (Klein et al., 2022).

Conclusion

The projects undertaken in this dissertation explore how interactions among biological entities result in emergent properties across varied spatiotemporal and organizational scales. Even though most of my projects focus on a specific problem in microbial ecology, they develop new general tools that can help answer similar questions in ecology and evolution. These tools span a wide range of equation-based and rule-based modeling frameworks, in addition to novel ways of applying statistical learning methodologies to large data volumes generated from these complex system models. I hope that these tools and techniques can contribute to the rising field of biological complex systems.

References

1. Auyang, S.Y., 1998. Foundations of complex-system theories: in economics
2. Chesson, P., 2000. Mechanisms of maintenance of species diversity. *Annual review of Ecology and Systematics*, 31(1), pp.343-366.
3. Eigenbrode, J.L. and Freeman, K.H., 2006. Late Archean rise of aerobic microbial ecosystems. *Proceedings of the National Academy of Sciences*, 103(43), pp.15759-15764.
4. Farquhar, J., Cliff, J., Zerkle, A.L., Kamyshny, A., Poulton, S.W., Claire, M., Adams, D. and Harms, B., 2013. Pathways for Neoproterozoic pyrite formation constrained by mass-independent sulfur isotopes. *Proceedings of the National Academy of Sciences*, 110(44), pp.17638-17643.
5. Faust, K. and Raes, J., 2012. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8), pp.538-550.
6. Fuhrmann, J.J., 2021. Microbial metabolism. In *Principles and applications of soil microbiology* (pp. 57-87). Elsevier., evolutionary biology, and statistical physics. Cambridge University Press.
7. Golding, S.D. and Glikson, M. eds., 2010. *Earliest life on Earth: habitats, environments and methods of detection*. Springer Science & Business Media.
8. Gosak, M., Markovič, R., Dolenšek, J., Rupnik, M.S., Marhl, M., Stožer, A. and Perc, M., 2018. Network science of biological systems at different scales: A review. *Physics of life reviews*, 24, pp.118-135.

9. Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.H., Weiner, J., Wiegand, T. and DeAngelis, D.L., 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science*, 310(5750), pp.987-991.
10. Hoel, E.P., Albantakis, L. and Tononi, G., 2013. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49), pp.19790-19795.
11. Klein, B., Hoel, E., Swain, A., Griebenow, R. and Levin, M., 2021. Evolution and emergence: higher order information structure in protein interactomes across the tree of life. *Integrative Biology*, 13(12), pp.283-294.
12. Hutchinson, G.E., 1961. The paradox of the plankton. *The American Naturalist*, 95(882), pp.137-145.
13. Izon, G., Zerkle, A.L., Zhelezinskaia, I., Farquhar, J., Newton, R.J., Poulton, S.W., Eigenbrode, J.L. and Claire, M.W., 2015. Multiple oscillations in Neoarchean atmospheric chemistry. *Earth and Planetary Science Letters*, 431, pp.264-273.
14. Izon, G., Zerkle, A.L., Williford, K.H., Farquhar, J., Poulton, S.W. and Claire, M.W., 2017. Biological regulation of atmospheric chemistry en route to planetary oxygenation. *Proceedings of the National Academy of Sciences*, 114(13), pp.E2571-E2579.
15. Kareva, I., 2011. Prisoner's dilemma in cancer metabolism. *PloS one*, 6(12), p.e28576.
16. Kelsic, E.D., Zhao, J., Vetsigian, K. and Kishony, R., 2015. Counteraction of antibiotic production and degradation stabilizes microbial communities. *Nature*, 521(7553), pp.516-519.

17. Kerr, B., Riley, M.A., Feldman, M.W. and Bohannan, B.J., 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature*, 418(6894), pp.171-174.
18. Kirkup, B.C. and Riley, M.A., 2004. Antibiotic-mediated antagonism leads to a bacterial game of rock–paper–scissors in vivo. *Nature*, 428(6981), pp.412-414.
19. Klein, B. and Hoel, E., 2020. The emergence of informative higher scales in complex networks. *Complexity*, 2020.
20. Klein, B., Swain, A., Byrum, T., Scarpino, S.V. and Fagan, W.F., 2022. Exploring noise, degeneracy, and determinism in biological networks with the einet package. *Methods in Ecology and Evolution*. <https://doi.org/10.1111/2041-210X.13805>
21. Ladyman, J., Lambert, J. and Wiesner, K., 2013. What is a complex system?. *European Journal for Philosophy of Science*, 3(1), pp.33-67.
22. López-García, P. and Moreira, D., 2020. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nature microbiology*, 5(5), pp.655-667.
23. McLane, A.J., Semeniuk, C., McDermid, G.J. and Marceau, D.J., 2011. The role of agent-based models in wildlife ecology and management. *Ecological modelling*, 222(8), pp.1544-1556.
24. Morrison, M., 2015. Why is more different?. In *Why More Is Different* (pp. 91-114). Springer, Berlin, Heidelberg.
25. Muller, E.E., Faust, K., Widder, S., Herold, M., Arbas, S.M. and Wilmes, P., 2018. Using metabolic networks to resolve ecological properties of microbiomes. *Current Opinion in Systems Biology*, 8, pp.73-80.

26. Orth, J.D., Thiele, I. and Palsson, B.Ø., 2010. What is flux balance analysis?. *Nature biotechnology*, 28(3), pp.245-248.
27. Peterson, S.B., Bertolli, S.K. and Mougous, J.D., 2020. The central role of interbacterial antagonism in bacterial life. *Current Biology*, 30(19), pp.R1203-R1214.
28. Pfeiffer, T., Schuster, S. and Bonhoeffer, S., 2001. Cooperation and competition in the evolution of ATP-producing pathways. *Science*, 292(5516), pp.504-507.
29. Regenmortel, M.H.V., 2004. Reductionism and complexity in molecular biology: Scientists now have the tools to unravel biological complexity and overcome the limitations of reductionism. *EMBO reports*, 5(11), pp.1016-1020.
30. Rothman, D.H., Fournier, G.P., French, K.L., Alm, E.J., Boyle, E.A., Cao, C. and Summons, R.E., 2014. Methanogenic burst in the end-Permian carbon cycle. *Proceedings of the National Academy of Sciences*, 111(15), pp.5462-5467.
31. Rousk, J. and Bengtson, P., 2014. Microbial regulation of global biogeochemical cycles. *Frontiers in microbiology*, 5, p.103.
32. Seth, E.C. and Taga, M.E., 2014. Nutrient cross-feeding in the microbial world. *Frontiers in microbiology*, 5, p.350.
33. Schuster, S., Boley, D., Möller, P., Stark, H. and Kaleta, C., 2015. Mathematical models for explaining the Warburg effect: a review focussed on ATP and biomass production. *Biochemical Society Transactions*, 43(6), pp.1187-1194.
34. Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R. and Ruppin, E., 2011. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS computational biology*, 7(3), p.e1002018.

35. Smets, B.F. and Barkay, T., 2005. Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. *Nature Reviews Microbiology*, 3(9).
36. Swain, A. and Fagan, W.F., 2019. A mathematical model of the Warburg Effect: Effects of cell size, shape and substrate availability on growth and metabolism in bacteria. *Mathematical Biosciences and Engineering*, 16(1).
37. Swain, A., Kaufman, A.J., Kalinowski, M., Yarwood, S.A. and Fagan, W.F., 2021a. Were Neoproterozoic atmospheric methane hazes and early Paleoproterozoic glaciations driven by the rise of oxygen in surface environments?. *bioRxiv*, pp.2020-10.
38. Swain, A., Hoffman, T., Leyba, K. and Fagan, W.F., 2021b. Exploring the evolution of perception: An agent-based approach. *Frontiers in Ecology and Evolution*, p.457.
39. Swain, A., Fussell, L. and Fagan, W.F., 2022. Higher-order effects, continuous species interactions, and trait evolution shape microbial spatial dynamics. *Proceedings of the National Academy of Sciences*, 119(1).
40. Szolnoki, A., Mobilia, M., Jiang, L.L., Szczesny, B., Rucklidge, A.M. and Perc, M., 2014. Cyclic dominance in evolutionary games: a review. *Journal of the Royal Society Interface*, 11(100), p.20140735.
41. Tarnita, C.E., 2018. Fast evolution unlocks forbidden communities. *Nature ecology & evolution*, 2(10), pp.1525-1526.
42. Thurner, S., Hanel, R. and Klimek, P., 2018. Introduction to the theory of complex systems. Oxford University Press.
43. von Bronk, B., Götz, A. and Opitz, M., 2018. Complex microbial systems across different levels of description. *Physical biology*, 15(5), p.051002.

44. West, S.A., Griffin, A.S., Gardner, A. and Diggle, S.P., 2006. Social evolution theory for microorganisms. *Nature reviews microbiology*, 4(8), pp.597-607.
45. Williford, K.H., Ushikubo, T., Lepot, K., Kitajima, K., Hallmann, C., Spicuzza, M.J., Kozdon, R., Eigenbrode, J.L., Summons, R.E. and Valley, J.W., 2016. Carbon and sulfur isotopic signatures of ancient life and environment at the microbial scale: Neoarchean shales and carbonates. *Geobiology*, 14(2), pp.105-128.
46. Zerkle, A.L., Claire, M.W., Domagal-Goldman, S.D., Farquhar, J. and Poulton, S.W., 2012. A bistable organic-rich atmosphere on the Neoarchaeon Earth. *Nature Geoscience*, 5(5), pp.359-363.
47. Zhuang, K., Vemuri, G.N. and Mahadevan, R., 2011. Economics of membrane occupancy and respiration-fermentation. *Molecular systems biology*, 7(1), p.500.

Chapter 2 : A mathematical model of the Warburg Effect: Effects of cell size, shape and substrate availability on growth and metabolism in bacteria

Please note that the contents of this chapter have been published as:

Swain, A. and Fagan, W.F., 2019. A mathematical model of the Warburg Effect: Effects of cell size, shape and substrate availability on growth and metabolism in bacteria. *Mathematical Biosciences and Engineering*, 16(1).

DOI: <https://doi.org/10.3934/mbe.2019009>

Abstract

The Warburg effect refers to a curious behavior observed in many organisms and cell types including cancer cells, yeast, and bacteria, wherein both the efficient aerobic respiration pathway and the inefficient fermentation pathway are utilized, despite the presence of ample oxygen. Also termed as overflow metabolism in bacteria, this phenomenon has remained an enigmatic and poorly understood phenomenon despite years of experimental work. Here, we focus on bacterial cells and build a model of three tradeoffs involved in the utilization of aerobic and anaerobic respiration pathways (rate versus yield, surface area versus volume, and fast versus slow biomass production) to explain the observed behavior in cellular systems. The model so constructed also predicts changes in the relative usage of both pathways in terms of size and shape constraints of the cell and identifies how substrate availability influences growth rate. Additionally, we use the model to explain certain complex phenomena in modern- and paleo-ecosystems, via the concept of overflow metabolism.

Introduction

Nutrients and substrates play a pivotal role in the sustenance and proliferation of living organisms, and therefore, evolutionary forces have constrained organisms to detect their availability. Cells reproduce quickly in a nutrient and substrate-rich environment and reduce proliferation when those resources become scarce. Availability of both energy rich substrates and starting materials for biomass production control this balance of growth and starvation (Vander Heiden et al., 2009).

Energy is produced in heterotrophic organisms by degrading organic compounds into products with low free energy. As per the second law of thermodynamics, no process can be fully efficient. So, a part of the free energy is used up to drive the reaction itself and the rest is stored into ATPs, the cell's energy currency. When more than one process is available for such conversions, there usually occurs a trade-off between rate (kinetic efficiency) versus yield (thermodynamic efficiency). This is true for fermentation and oxidative phosphorylation (OP) (Schuster et al., 2015).

Fermentation yields 2-4 moles of ATP per mole of glucose whereas OP produces about 30 ATPs (the actual number may vary with the organism). In contrast, the rate of fermentation can reach about two orders of magnitude higher than the rate of OP (Voet and Voet, 2004; Schuster et al., 2015).

Many organisms can utilize both pathways in the presence of oxygen and substrate (Pfeiffer et al., 2001), but OP is not feasible in the absence of oxygen. A direct consequence of this is seen in a phenomenon, first observed in cancer cells (Warburg, 1956), called the Warburg effect, where cells utilize the fermentation pathways much more than the OP in presence of ample substrate and even oxygen (Schuster et al., 2015). A similar phenomenon can be seen across a broad range of

cell types and organisms. Many yeast variants (in which the same phenomenon is called the Crabtree effect (Deken, 1966)) and bacteria (where it is called overflow metabolism (Basan et al., 2015)) exhibit this behavior. Other than the cells undergoing growth, lymphocytes and Kupffer cells (upon activation), fast twitch muscle fibers and microglial cells, also display this effect in the human body (Schuster et al., 2015; Vander Heiden et al., 2009). Most of the work on the Warburg effect has been concentrated on cancer cells, and for bacterial metabolism, the phenomenon has been well studied in *Escherichia coli* (Zhuang et al., 2011).

This behavior can appear to be paradoxical in that a less efficient pathway is utilized even when a more efficient one is available. Given that substrate limitation is a major environmental constraint, the properties of these two pathways must have been under intense selection. Such a behavior, therefore, should confer some additional biological advantage, and a number of explanations and modeling frameworks have been suggested accordingly (Schuster et al., 2015).

Several game theory models have been proposed describing a population using a common resource (glucose in most cases), and the two metabolic pathways (i.e. OP and fermentation) as alternate strategies based on a kinetic-thermodynamic trade-off. A tragedy of the commons scenario sometimes emerges (Pfeiffer et al., 2001; Kareva. 2011). Prisoner's dilemma and Nash equilibria (based on relative enzymatic/protein costs of the two pathways) were also used to describe the dynamics of this system (Frick et al., 2003; Schuster et al., 2015).

For a similar trade-off between investment in protein synthesis and metabolic yield, a medium-scale kinetic model has also been proposed (Molenaar et al., 2009). Such a model explains the increase in ribosomal content observed with increasing growth rate and related characteristics in bacteria (Schuster et al., 2015).

Another family of models that has been used extensively to explain this behavior belongs to a framework called Flux Balance analysis (FBA) (Famili et al., 2003; Price et al., 2004; Rupp et al., 2010). In such an analysis, linear maximization is utilized to determine optimal values of reaction rates at a steady state condition where a linear combination of rates attain a maximized value. There has been some controversy on whether the criterion for optimality should be based on rate or yield in these models (Schuster et al., 2015). A branch-off from this family of models, called FBA with macromolecular crowding (FBAwmc), incorporates the areal and volumetric needs of the enzymes constrained by the cell membrane space and cytoplasm, respectively (Zhuang et al., 2011). Some studies looked at maximization of biomass production using this framework (Shlomi et al., 2011).

Some of the models discussed above focus on one specific behavior seen in the Warburg effect but avoid other behaviors that are part of the same process, thus preventing study of interactions among the observed behaviors. Among notable ones, Pfeiffer et al., (2001) use the rate-efficiency trade-off to explore the problem and delve into origins of multicellularity. Kareva (2011) uses efficiency of the two pathways (respiration and fermentation) and the effect of lactic acid produced from fermenting cells, on neighbouring cells but it looks at cancer cells in specific.

In many other cases, the work utilizes more than one trade-offs, especially using the FBA regime (Beg et al., 2007; Vazquez et al., 2010; Shlomi et al., 2011; Zhuang et al., 2011, Schuster et al., 2015). Vazquez et al., (2010) focuses on cancer cells, and uses constraints on maximal glucose uptake, enzyme efficiency and total cellular volume, in order to maximize ATP production. Their model considers not only the stoichiometry of glycolysis and oxidative phosphorylation but also the enzyme-volumetric costs of activating these pathways. A comparable idea of constraints on enzyme-volumetric costs was used in a work focusing on *Escherichia coli*

to explain overflow metabolism (Beg et al., 2007). Yet another work which used a similar concept was that of Shlomi et al. (2011), which focuses on cancer cells in specific. For this class of models, a consistent criticism has been that, despite its overwhelming success, these models fail to explain overflow metabolism from first principles, that is, without imposing multiple measured fluxes as constraints on the models (Molenaar et al., 2009). In comparison, a slightly different approach was taken in the work by Molenaar et al., (2009), where they focus on trade-offs between investments in enzyme synthesis and metabolic yields for different catabolic pathways, in order to maximize growth rate. Zhuang et al., (2011) uses the constraint on membrane occupancy to explore the effects on fermentation and respiration, via a transformed FBA.

Given this large body of work on modeling the Warburg effect, yet another model may seem to be a futile exercise, but if one looks closely, most of the models either focus on cancer cells or dive only into the biochemical aspect of the problem. This leaves the ecological implications of the Warburg effect to be modeled, specifically in bacteria. None of the models discussed above take into account the effects associated with size and shape of the bacterial cell on the overall metabolism of the organism, especially considering the wide variability in both those physical characteristics. Moreover, predicting overflow metabolism from first principles has always been a challenge without invoking a large number of boundary conditions (Molenaar et al., 2009).

In this work, we attend to the three major trade-offs involved in the Warburg effect, specifically in the case of a bacterial cell. We seek to incorporate only the necessary details, both from a physical and a biochemical viewpoint, to retain model simplicity and usability in an ecological context (See **Table 2.1**).

Our aim here is to predict size and shape dependence of growth rate in bacteria across a range of substrate concentrations, in addition to estimating the proportions of OP and fermentation (hereafter denoted AG, for aerobic glycolysis in presence of oxygen) for respiration and metabolism. We also seek to explain overflow metabolism in our model as a direct consequence of energy production and consumption trade-off, through first principles. Unlike the previous models on the Warburg effect, we explore more of the ecological ramifications through the model, in addition to the basic biochemical effects. In turn, we limit ourselves to a single bacterial cell as the model system and glucose as the substrate.

Trade-offs involved in a bacterial cell in the Warburg Effect	Seen in
1. Yield versus rate: Fermentation yields about 2-4 moles of ATP per mole of glucose whereas OP produces about 30 ATPs. However, rates of fermentation can reach about two orders of magnitude higher than that of OP (Voet and Voet, 2004; Schuster et al., 2015)	Game theory models (Pfeiffer et al., 2001; Kareva, 2011)
2. Surface area versus volume: The enzymes responsible for OP are located on the cell surface whereas those for fermentation are in the cytoplasm	FBAwmc (Zhuang et al., 2011)
3. Biomass production trade-off: Fermentation facilitates biomass generation (Vander Heiden et al., 2009)	FBA and related models (Shlomi et al., 2011)

Table 2.1: Trade-offs involved in Warburg effect

The Model

The model assumes a bacterial cell, with glucose transporter proteins (responsible for importing glucose into the cell) and cytochrome oxidase enzymes (where OP occurs) embedded in the cell membrane as trans-membrane proteins (Zhuang et al., 2011). In contrast, the (upper) glycolytic enzymes, which convert glucose to pyruvate, and the fermentation enzyme pathways function in the cytoplasm (Schuster et al., 2015). Thus, OP should scale as a function of the surface area of the cell whereas fermentation processes should scale as a function of cell volume. A schematic representation of the model is illustrated in **Figure 2.1**.

The rate of glucose input into the cell (r_0) can be represented as,

$$r_0(t) = \frac{d G_{cell}(t)}{d t} = \beta(t) \cdot A(a_{cell}) \cdot k_{\beta} \cdot G_0(r, t) \quad \dots (1)$$

where t is time, r is position, G_{cell} is the cellular concentration glucose, G_0 is the concentration of glucose in the environment around the cell, β is the areal density of glucose transporter sites on the cell membrane, A is area of the cell membrane (as a function of the semi-major axis of the cell, a_{cell}), and k_{β} is normalized efficiency of the transporters (we define normalized efficiency as the ratio of molecules transported relative to the total number present on the outer/source side per unit time). We assume a uniform G_0 in the environment around the cell here and therefore, inside it (G_{cell}).

This glucose imported into the cell goes into the upper glycolysis pathway, whose enzymes are located in the bacterial cytoplasm. In bacteria, there are two well-known pathways for this: Embden-Meyerhof-Parnas (EMP) pathway and Entner-Doudoroff (ED) pathway, which differ in both their output of ATP (2 mol in EMP versus 1 mol in ED for every mol of glucose consumed)

and their enzyme costs (the EMP enzyme costs can be 3.5 to 5 fold higher than ED pathway enzymes) (Flamholtz et al., 2013). The end product is pyruvate in both the cases, and the rate of conversion (r_1) can be written as:

$$r_1(t) = (\delta_{EMP} \cdot \kappa_{EMP}(t) \cdot k_{EMP} + \delta_{ED} \cdot \kappa_{ED}(t) \cdot k_{ED}) \cdot V(a_{cell}) \cdot r_0(t) \quad \dots (2)$$

where κ_{EMP} and κ_{ED} are the volume density of EMP and ED enzyme sites respectively, k_{EMP} and k_{ED} are normalized efficiencies of the pathways, and V is the volume of the cell. δ_{EMP} and δ_{ED} indicate presence of EMP and/or ED pathway in the organism under scrutiny, taking a value 1 if present and 0 if absent; else if both are present, it denotes the fraction out of 1 which is occupied by either pathway such that $\delta_{EMP} + \delta_{ED} = 1$. About 43% of the prokaryotic cells use just EMP pathway, 13% use just the ED pathway, 14% use both pathways, and in about 30% of the cases, the pathway is not known (Flamholtz et al., 2013).

The relative usage of the EMP and ED pathways can be calculated on the basis of protein cost, kinetics and the efficiency of the enzymes involved in both the pathways at any given glucose concentration as well as the thermodynamic constraints (Flamholtz et al., 2013). To put it in simpler terms, the ED pathway is about 3-5 times faster than EMP pathway, but EMP produces 2 ATP molecules whereas ED produces 1 per molecule of glucose (Flamholtz et al., 2013). So, in the case of OP, a difference of one ATP is not of much significance because the total production of ATP is high, but in cases of fermentation, one ATP is a substantial difference.

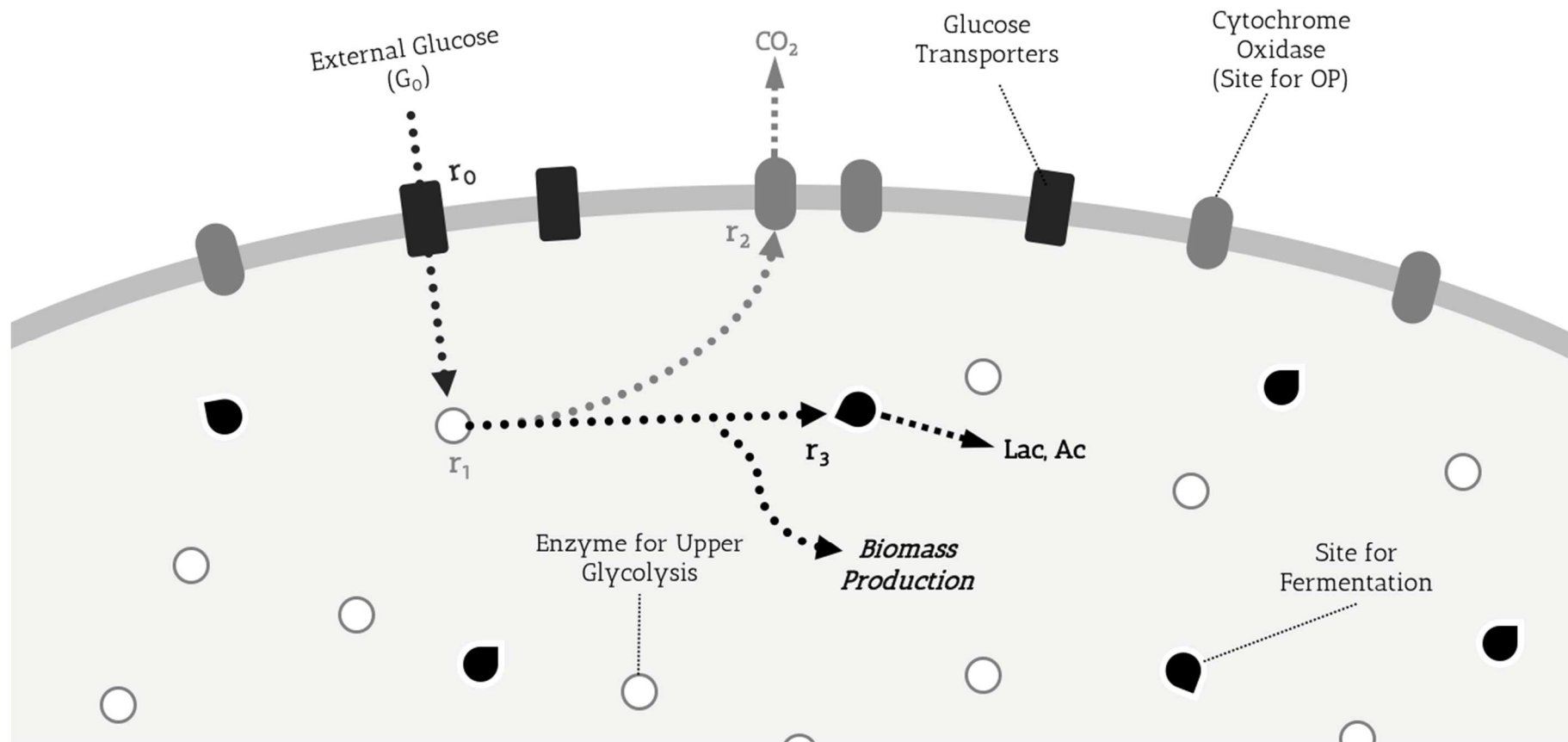


Figure 2.1: A simplified representation to model Warburg effect. The external glucose is imported into the bacterial cell via glucose transporters located on the cell membrane at a rate r_0 . This glucose is utilized by upper glycolysis pathways to convert it to pyruvate at a rate r_1 . This converted substrate is then divided into two paths: either to OP (at a rate r_2 at the cytochrome oxidase located in the membrane) or to fermentation/aerobic glycolysis (AG) (at a rate r_3 at the fermentation enzyme sites located in the cytoplasm). Note that only the substrate used in the latter shunt goes into biomass production whereas substrate used in OP is lost as CO_2 .

Thus, we would expect to see mostly EMP in anaerobes versus ED in aerobes. Anaerobic bacteria have predominantly only EMP (97%), whereas some facultative anaerobes and aerobes have ED only (10% and 21% respectively), others have both ED and EMP (19% and 21% respectively), and more than half have only EMP (Flamholtz et al., 2013). Here we seek to model aerobic heterotrophic bacteria, so we can expect that the ED pathway is prevalent when the metabolism is predominantly aerobic (OP) while EMP is prevalent when it is anaerobic (fermentation).

Next, this $r_1(t)$ flux of products from upper glycolysis goes to either the oxidative phosphorylation (OP) pathway ($r_2(t)$) or to the fermentation pathway ($r_3(t)$). This splitting of the flux forms the basis of the three trade-offs on which we focus: yield vs rate, surface area vs volume, and biomass production (**Table 2.1**). The rates for r_2 and r_3 , can be written as:

$$r_2(t) = (\gamma_{Cyo} \cdot k_{Cyo} + \gamma_{CydII} \cdot k_{CydII}) \cdot A(a_{cell}) \cdot r_1(t) \quad \dots (3)$$

and

$$r_3(t) = \sigma \cdot k_{\sigma} \cdot V(a_{cell}) \cdot r_1(t) \quad \dots (4)$$

where γ_{Cyo} and γ_{CydII} are the areal density of the two cytochrome complexes Cyo and Cyd-II respectively, k_{Cyo} and k_{CydII} are normalized efficiency of the complexes, σ is the volume density of the fermentation enzymes (sites), and k_{σ} is the normalized efficiency of the fermentation enzymes (sites). Here, normalized efficiency is defined as the ratio of products formed by the enzyme site relative to the total substrate that was present, per unit time. As a part of the simplification in the model, we have only considered the cytochrome complexes Cyo and Cyd-II as a part of the model leaving behind the Cyd-I complex.

For maintenance of the cellular membrane structure in prokaryotes, the protein-to-lipid ratio is kept within certain bounds (Molenaar et al., 2009). At high rates of catabolism, the surface

area available to proteins can become saturated and proteins might need to 'compete' for expression (Zhuang et al., 2011). Both substrate transporters (in this case, glucose transporters) and OP enzymes are located on the membrane, and thus, compete for the available space for expression. Thus, we can write:

$$r_2(t) \leq r_{2,max} \quad \dots (5)$$

where, we assume a constant maximal value.

In previous literature, the relative membrane cost of an enzyme has been inversely related to its turnover rate (Bekker, et al. 2009) and hence at high catabolic rate one would expect the faster and inefficient Cyd-II to be preferred over efficient but slower Cyd (Bekker et al., 2009). The relative cost of moderately efficient Cyd-I is similar to Cyo under completely aerobic conditions, but under low concentration of oxygen, the cost becomes much less due to the high affinity of Cyd-I to O_2 (Tseng et al., 1996; Zhuang et al., 2011).

Summarizing the model, the net energy production rate can be written as

$$e_+(t) = \frac{dE_{gain}}{dt} = e_{r_2}(t) + e_{r_3}(t) = \langle \epsilon_{OP} \rangle(t) \cdot r_2(t) + \langle \epsilon_F \rangle(t) \cdot r_3(t) \quad \dots (6)$$

where e_{r_2} is the energy produced through r_2 (via OP) and e_{r_3} is the energy produced through r_3 (via fermentation (F) or in presence of oxygen, termed as aerobic glycolysis). $\langle \epsilon_{OP} \rangle$ and $\langle \epsilon_{OP} \rangle(t)$ are the weighted average efficiencies of ATP production from OP (via Cyo and Cyd-II), and F respectively. For simplicity, each of these process averages also include a term regarding the overall efficacy of the upper glycolysis (as a function of the proportion of EMP and ED pathways). In other words, the two averages mentioned above involve the efficiency from the point of upper glycolysis to the end of their respective processes.

A main function of unregulated glycolysis in proliferating cells is to maintain the supply of intermediates required to support biosynthesis (Lunt et al., 2011; Vander Heiden et al., 2009). Interestingly, in such rapidly proliferating cells, most of the glucose is converted into lactate (or other fermentation products) and excreted. However, an increased lactate excretion is correlated with increased growth (Lunt et al., 2011; Vander Heiden et al., 2009; Basan et al., 2015). Such a seemingly inefficient high glycolytic flux to fermentation products and low flux to biosynthesis actually helps in regulation of biomass production during growth (Lunt et al., 2011; Vander Heiden et al., 2009). The branching of a low flux process (biosynthesis) from a high flux one (fermentation) would render the former extremely sensitive to the latter (Lunt et al., 2011; Newsholme et al., 1985). Comprehending the exact role of such processes in biomass production would need a thorough study of carbon metabolism (for details please see Lunt et al., 2011). However, given the dependence, one can certainly assume a simple linear relationship between biomass synthesis and fermentation.

Here onwards, we explain biomass production in terms of r_3 flux and thus, the rate of biomass addition Λ , can be given by:

$$\Lambda(t) = g \cdot r_3(t) \quad \dots (7)$$

where g is the conversion constant relating the products of fermentation pathway to biomass. The process of biomass production also involves energy usage (along with the need for other resources such as nitrogen and phosphorus, which have been assumed to be present in sufficient quantity for this model).

The energy consumption rate can thus, be given by:

$$e_-(t) = e_b + \epsilon_g \cdot g \cdot r_3(t) \quad \dots (8)$$

where e_b is the basal energy requirement of the cell and ϵ_g is the energy required per unit conversion for biomass production from base materials.

Most metabolic networks operate in a steady state (Schuster et al., 2015). Thus, here we tried to optimize the energy production and growth using linear programming, adhering to the following relations:

$$r_0(t) \geq r_2(t) + r_3(t) \quad \dots (9)$$

$$e_b \leq \langle \epsilon_{OP} \rangle(t) \cdot r_2(t) + \langle \epsilon_F \rangle(t) \cdot r_3(t) \quad \dots (10)$$

$$\frac{r_2}{(\gamma_{Cyo} \cdot k_{Cyo} + \gamma_{CydII} \cdot k_{CydII}) \cdot A(a_{cell})} - \frac{r_3}{\sigma \cdot k_\sigma \cdot V(a_{cell})} = 0 \quad \dots (11)$$

The first inequality states that the total glucose processing that takes place in OP and fermentation is constrained by the total glucose input into the cell. The next relation describes the fact that the total energy production in the cell has a lower bound of the basal metabolic rate of the cell (e_b). Below this level, the cell may cease to function properly or shut down entirely (Kempes et al., 2017). The third relation is an equation which is modified and combined form of equations (3) and (4). The way to look at it is that both the terms in Equation (10) simplify to r_1 , as is evident from the equations (3) and (4).

Let us also define an additional parameter called Z denoting the ratio of surface area-to-volume of a cell. From (11), it is evident that Z controls the distribution of r_1 into r_2 and r_3 by controlling the total number of sites available for those processes. A low value of Z denotes either a spherical cell and/or a large cell size; and a high value of Z represents an elongated/deformed cell shape and/or a small cell size.

These conditions can be used for linear programming and the point in solution space having the highest value of r_3 will give the optimal values of (r_2, r_3) at a given set of steady state conditions. This accords with our assumption that only r_3 related processes directly help in biomass production, and any organism would maximize its biomass production in a given set of conditions, provided its basal needs are satisfied. All other parameters are considered to be constants for finding the optimal solution for a given iteration.

Results

Given the number of model parameters, it is pragmatic to focus on a few important ones, which would enable us to predict and investigate broad-scale natural phenomena from an improved viewpoint. In this work, we primarily focus on changing the substrate concentration and the surface area-to-volume ratio of bacterial cells to gain insights from the model and explore the Warburg effect (overflow metabolism) in bacterial cells.

Dependence on substrate concentration

r_0 is the rate of substrate input from the environment, and is representative of the external substrate concentration. **Figure 2.2** plots the results of the applying the maximization criterion for r_3 constrained by (9)-(11), on changing the value of r_0 . Three important regimes of substrate concentration are considered: glucose starved (low r_0), moderate glucose availability (moderate r_0), and glucose rich environment (high r_0).

When (9) and (10) do not have a common solution space, i.e. when the substrate availability is not sufficient to sustain all basic metabolic activities, but is high enough not to cause complete dormancy (Kempes et al., 2017), the maximized solution is $(0, r_0)$ in the (r_2, r_3) space. This

suggests that under a substrate-poor scenario there exists a need for high efficiency in substrate conversion, without much consideration for its rate of production or cellular growth.

Increasing r_0 enables high rates of r_3 due to fulfillment of basal metabolic processes and a need to increase biomass output for growth. This points at the higher growth rates associated with nutrient and substrate rich environments (Vander Heiden et al., 2009). At extremely high glucose concentrations, the value of β would saturate (to its highest value in order to keep the protein-to-lipid ratio of the membrane intact) and k_β would decrease (as only a specific maximal number of glucose molecules can be transported into the cell constrained by the protein efficiency) in Equation (1). These two features make the value of r_0 constant above a certain threshold of G_0 . Thus, beyond this threshold value, growth (which is proportional to r_3) would not depend upon G_0 explicitly, but only on the available r_0 . Such a behavior is observed in *E. coli* along with an increase in expected growth rate with increasing G_0 until the threshold concentration (maximal value of r_0) is reached (Shehata et al., 1971).

Such an overall pattern of starvation metabolism in environments with scarce substrate environments and proliferative metabolism in environments with abundant substrate is seen in all unicellular organisms (Vander Heiden et al., 2009).

Dependence on cellular size and shape

The solutions of the constraint equations (9)-(11) for three differing regimes of Z are plotted in **Figure 2.3**. The value of r_0 has been kept constant in the three cases for clarity. As mentioned previously, a low value of Z refers to either a spherical cell and/or a large cell size; and a high value of Z points to an elongated/deformed cell shape and/or a small cell size.

Large values of Z enable higher efficiency for energy conversion processes but trade off for slower growth, which may be typical for substrate poor environments. Low values of Z sacrifice efficiency for a higher rate of energy and biomass production and would be expected in conditions where substrate is abundant. This shape and size based control of metabolic strategies is quite commonly seen in different types of bacteria (Young, 2006). Numerous previous works and experiments support the predictions from our model regarding the interdependence of metabolism, growth, shape and size. These cases have been analyzed in the discussion section.

Overflow metabolism and bacterial metabolic regime

The Warburg effect is termed overflow metabolism in bacteria, due to the cellular excretion (or overflow) of excess metabolites like lactate, acetate, or ethanol after incomplete oxidation of

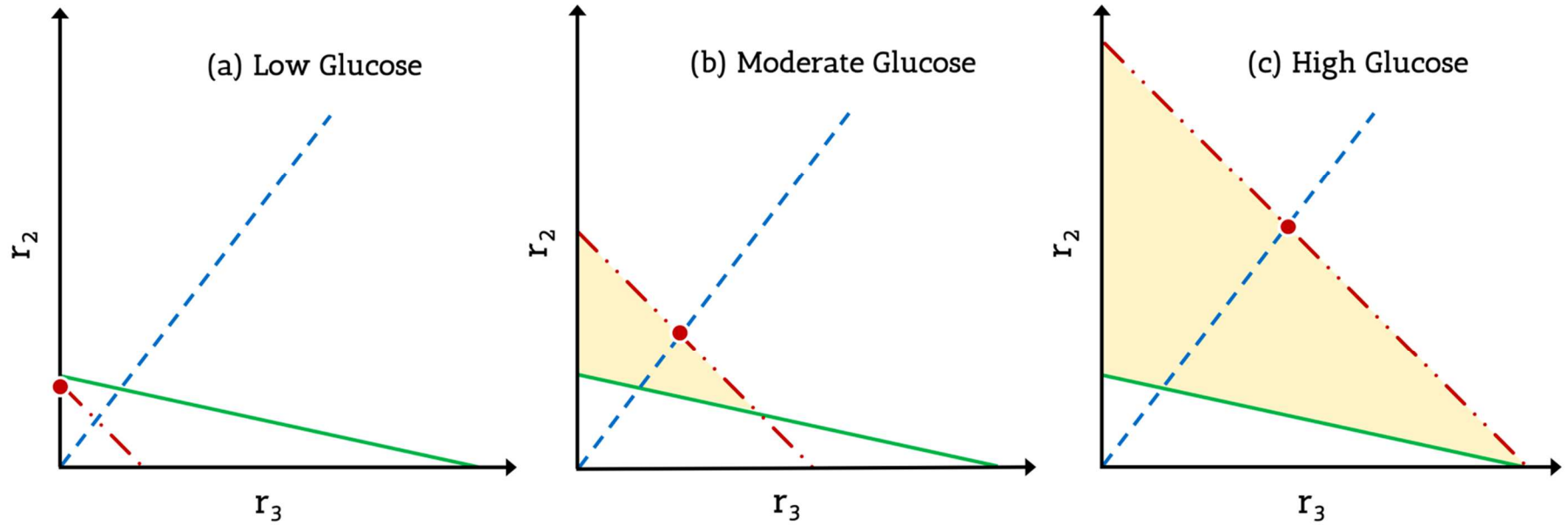


Figure 2.2. Representation of the solutions for a bacterial cell at steady state with all features except glucose concentration constant. The dot-dashed red line represents relation (9), the solid green line represents relation (10) and the dashed blue line represents Equation (11). The area shaded in yellow represents the solution for (9) and (10), and the red dot represents the optimized value of r_3 for (9)-(11).

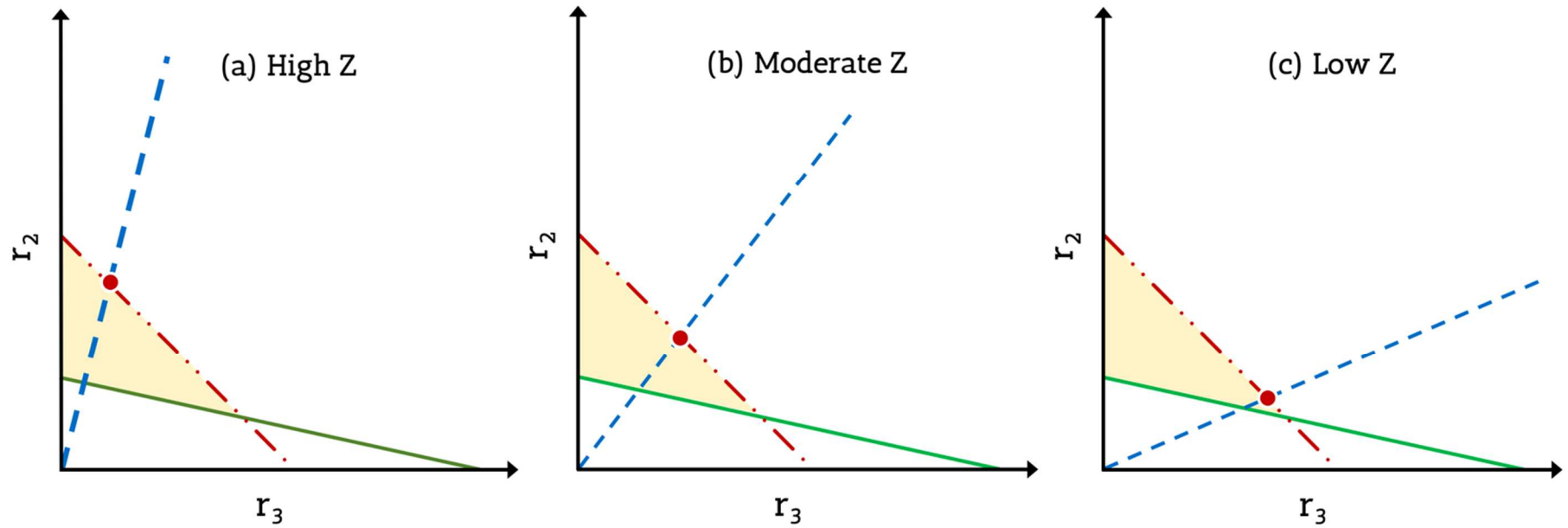


Figure 2.3. Representation of the solutions for a bacterial cell at steady state with all features constant except the ratio of surface area to volume of the cell (denoted by Z). The dot-dashed red line represents relation (9), the solid green line represents relation (10) and the dashed blue line represents Equation (11). The area shaded in yellow represents the solution for (9) and (10), and the red dot represents the optimized value of r_3 for (9)-(11).

glucose. Acetate overflow metabolism has been well documented in *E. coli* (Zhuang et al., 2011), but the process is not clearly understood (Vemuri et al., 2006).

During the process of conversion of r_3 products into biomass, energy consumption is involved as in Equation (8). If the value of $\epsilon_g \cdot g > \epsilon_F$, then the energy gained from r_3 alone is insufficient to convert substrate into biomass and the cell would depend upon more efficient r_2 for energy. But, if the value of r_3 is high enough e_+ (Equation 6) would be insufficient to sustain e_- (Equation 8), i.e. $e_- > e_+$. In such a case, the excess r_3 products would be dumped outside the cell and the rest would be used for biomass production. The maximum value of r_3 for which all substrate would be converted into biomass, is given by the relation:

$$\begin{aligned} e_+ \geq e_- &\Rightarrow \epsilon_2 \cdot r_2 + \epsilon_3 \cdot r_3 \geq e_b + \epsilon_g \cdot g \cdot r_3 \\ &\Rightarrow \epsilon_2 \cdot r_2 + (\epsilon_3 - \epsilon_g \cdot g) \cdot r_3 \geq e_b \end{aligned} \quad \dots (12)$$

The solution for (9) and (12) would give us the maximal value of r_3 beyond which export of excess r_3 products will occur. Additionally in such a high flux regime, we would need to invoke Equation (5), which describes the maximum rate of OP, in order not to destabilize the protein-lipid ratio of the bacterial membrane.

In such a scenario, the fermentation related enzymes are up-regulated (Lunt et al., 2011), resulting in changes to the slope of Equation (11), even though the value of Z remains unchanged. This occurs because the number of OP enzymes (γ) remaining constant whereas that of fermentation (σ) increasing in value. This decreases the slope of the line- a shift to match the intersection of Equations (5) and (9), i.e., in order to process the remaining glucose influx (shown in **Figure 2.4a** via the black dot, which is the final solution for Equations (5), (9)-(11)).

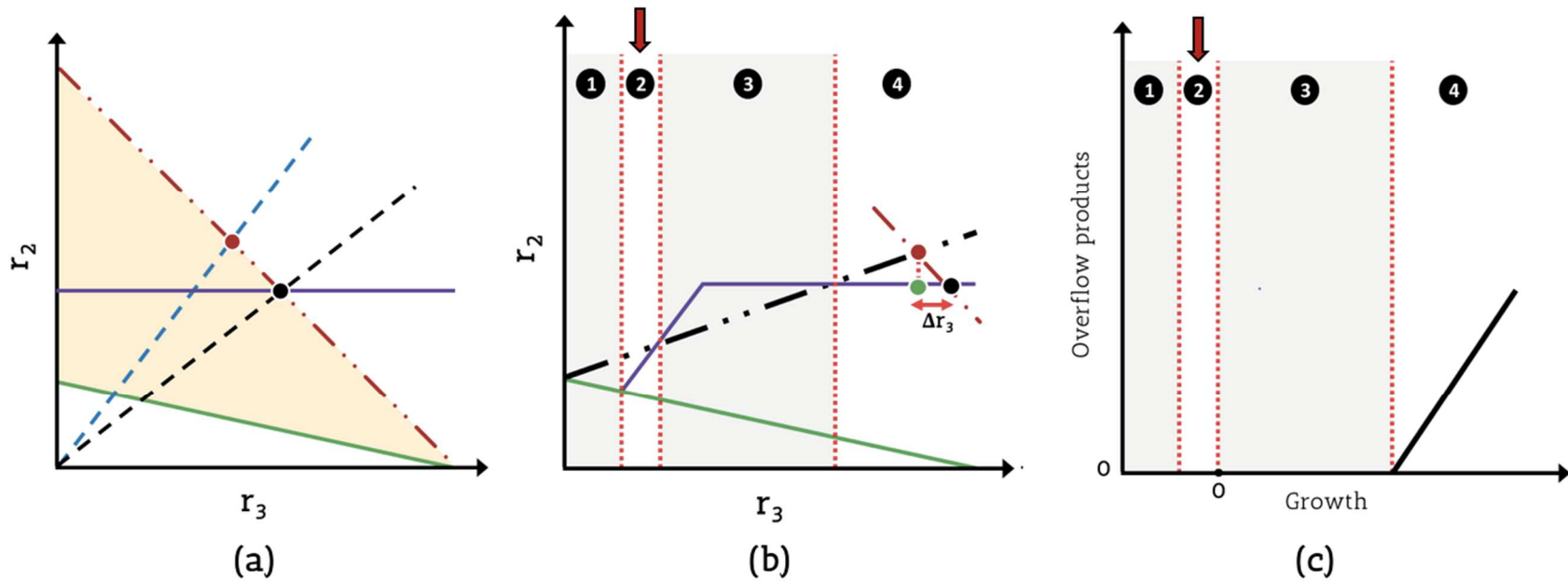


Figure 2.4. Representation of overflow metabolism in a bacterial cell. (a) The dot-dashed red line represents relation (9), the solid green line represents relation (10) and the dashed blue line represents Equation (11) and the solid purple line represents relation (5). The area shaded in yellow represents the solution for (9) and (10), and the red dot represents the optimized value of r_3 for (9)-(11). The black dot represents altered final solution due to (5). See text for details. (b) The solid green line represents (10), the solid purple line represents the final solutions of (5), (9)-(11) and the black dot dash line represents (12). Δr_3 represents the estimate of overflow metabolic products. The phase space has been divided in four parts based on the solid purple line and (12). (c) Overflow products vs. growth (Equation 7)

Furthermore, the intersections of the line representing final solutions of (5), (9)-(11) with the line representing (12), provides us with three different regimes of bacterial metabolism (denoted by regime 2-4 in **Figure 2.4b**). We also find an additional regime of metabolism arising from the part of the phase space where there is no solution for (5), (9)-(11) (denoted by regime 1 in **Figure 4b**). This leads us to a total of four such scenarios (as seen in **Figure 4b**). We discuss these regimes in brief before delving into overflow metabolism in detail again.

As discussed in section 3.1, regime 1 in **Figure 2.4b,c**, is characterized by substrate concentrations that are insufficient to fulfill all the basic metabolic requirements of the cell. Thus, the cell resorts to limited activity and uses only OP to extract as much energy as possible from the substrate for sustenance.

Regime 2 is more interesting and is denoted by an arrow in **figures 2.4b,c**. Here, the substrate flux is enough to maintain the basic cellular metabolic requirements, but does not produce enough energy for biomass production (i.e. $e_+ \geq e_b$ but $< \epsilon_g \cdot g \cdot r_3$). In such a case, even though it is a low substrate scenario, r_3 products are in excess. According to a previous model (Moller et al., 2018), in such a regime, excess r_3 products might be converted back to pyruvate r_1 product) and reused for OP to keep the respiration at full capacity. Unfortunately, a direct prediction of this could not be obtained from our model, but it predicts excess r_3 products that is definitely not known to be excreted away (see data in Basan et al., 2015).

Once enough energy can be produced for the cell to start proliferating (the first intersection of the solid purple line and the dot dashed black line in **Figure 2.4b**), the cell not only has enough energy to sustain basic metabolic activities but also for biomass production. In this case, there are no overflow products. This scenario continues till the second intersection of the aforementioned lines in **Figure 2.4b**. Above this limit, overflow metabolism takes place, because $e_+ \not\geq e_-$ and

thus the excess products are excreted. There are several other reasons for non-reduction of r_3 processes even though they seem wasteful and they can be found in relevant reviews (Lunt et al., 2011).

In **Figure 2.4b**, the difference (Δr_3) in the final solution of r_3 value for the solutions of (7), (9)-(11) with (12) will give an estimate of the overflow metabolism. In this fourth regime, Δr_3 can be plotted against growth/biomass production rate (Equation (7)) to obtain **Figure 2.4c**. The prediction of this relation matches numerous laboratory observations of overflow metabolism in multiple strains of bacteria (Basan et al., 2015).

Further considerations and applications

As represented in Equation (7), the bacterial growth rate R can be given by:

$$R = \frac{\Lambda}{\Lambda_0} = \frac{g \cdot r_{3,max}}{\Lambda_0} \quad \dots (13)$$

where $r_{3,max}$ is the final solution for (5), (9)-(12) (symbolized by the purple line in **Figure 2.4b**) and Λ_0 is the biomass (Λ) required to make a new cell. This growth rate, R , can then be incorporated into a regular Lotka-Volterra equation system to find the population size N_i of a species i directly as a function of substrate concentration through R , given by:

$$\frac{dN_i(t)}{dt} = R_i \cdot N_i(t) \left(1 - \frac{\sum_j \alpha_{ij} \cdot N_j(t)}{K_i} \right) \quad \dots (14)$$

where R_i is the growth rate of species i , N_i is the number of individuals of species i , α_{ij} is the competition coefficient between species i and j and K_i is the carrying capacity of species i .

Our model can be extrapolated to include multiple substrates using a vector to symbolize r_0 , in order to estimate R . We can then use the independent experimental data for growth and

substrate preference in individual species to calculate α_{ij} (Swain and Chatterjee, 2017), in order to obtain all the parameters in (13).

In addition to the direct estimation of populations, we can use the same technique of vectorization of resources and then species in the model to calculate the magnitude of overflow metabolism. This understanding can then be used in the context of many modern and paleo-ecosystems where spatio-temporal change in substrate concentration is significant and would result in interactions between various communities. Such usage is not possible from any of the earlier models because they do not estimate overflow metabolism from first-principles (Molenaar et al., 2009), and further, most the FBA models are specific to a particular species. Temperate wetlands are one such example of an ecosystem with spatio-temporal change in substrate concentration. Our model can be extrapolated to predict and estimate the gas fluxes (CO_2 and methane), and further to account for phenomena such as aerobic methanogenesis.

In temperate zone wetlands, the fall season can be seen as a pulse of high concentration of substrate and resources. Due to the Warburg effect, an increased concentration of substrates would not result in higher carbon dioxide fluxes, but rather in production of secondary metabolites like lactate and acetate, and a sudden proliferation of bacterial populations. This can not only affect the amount of carbon dioxide released but also the amount of substrate that is available to methanogens. Acetate is produced by heterotrophic bacteria through overflow metabolism (Basan et al., 2015) and can be utilized by the acetate-based methanogens, which form a large part of the methanogen community in the wetlands. Moreover, lactate is also formed during the process (Vander Heiden et al., 2009) and can be used by the syntrophic relations of hydrogen-based methanogens and sulfate-reducing bacteria (McInerney et al., 1981).

This methane produced by the methanogens, using the metabolites from the Warburg effect (overflow metabolism), is then oxidized by the methanotrophs, which use the oxygen from the water column (of the small water bodies in these wetlands) above the methanogens, thus producing carbon dioxide. We illustrate this idea in **Figure 2.5** where the bold red line is the total CO_2 production and the dotted red line represents the CO_2 emitted from the heterotrophic bacteria. So, even though the total CO_2 emissions increase, it is actually the increase in methane production and its subsequent oxidation that is fueling this rise.

Once the initial substrates start depleting, the water column becomes anoxic due to oxygen usage by both the methanotrophs, whose population rises because of increased methane availability, and by the already large aerobic heterotrophic community, fueled by the initial pulse of substrates. After a certain degree of anoxia is reached, the aerobic heterotrophs start declining due to lower oxygen and lower substrate availability. Only the anaerobic heterotrophs remain plentiful, but in reduced abundance as easily available resources have been depleted, and they have

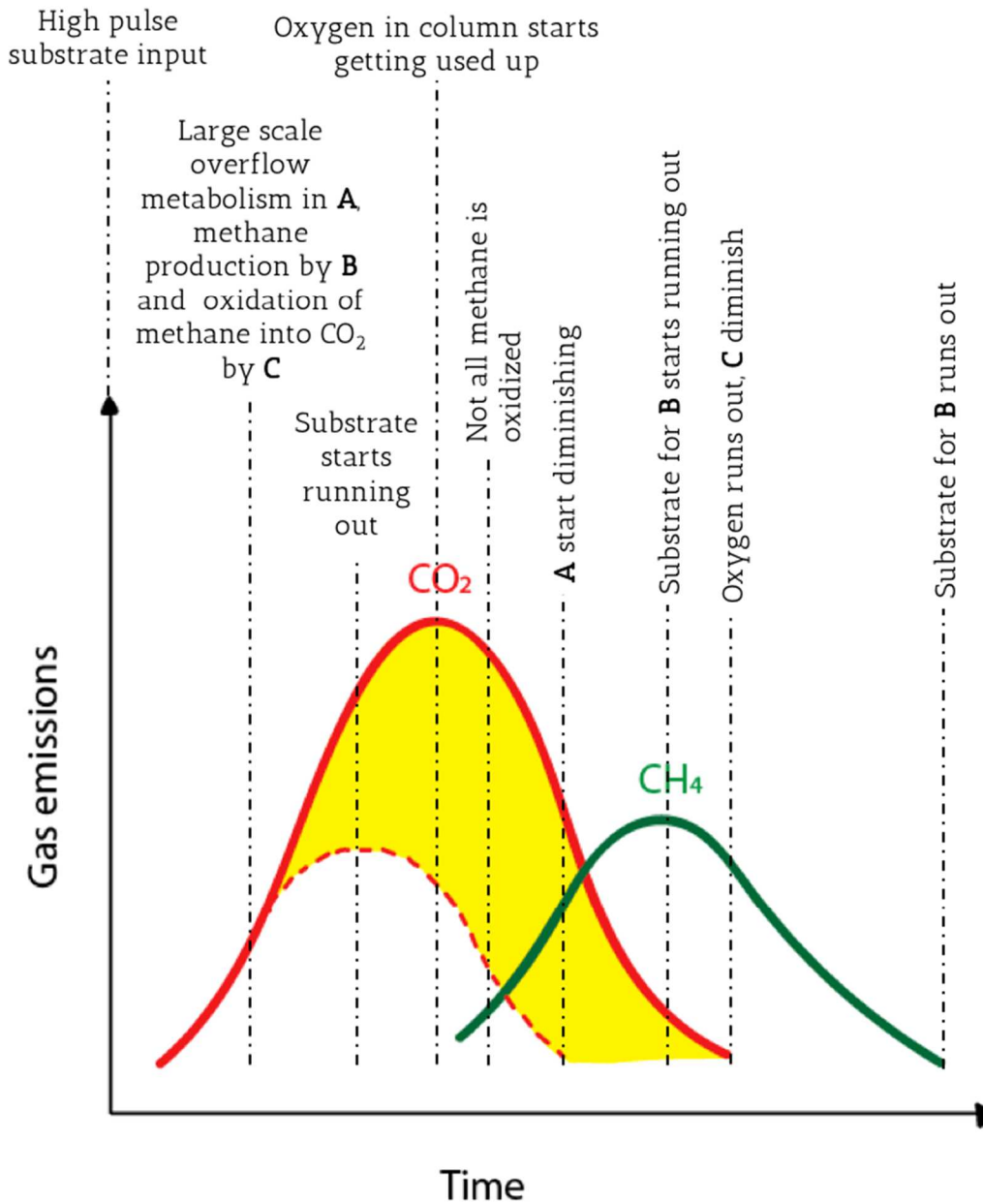


Figure 2.5. A representation of the effect of overflow metabolism in temperate zone wetlands. **A** refers to heterotrophic bacteria, **B** refers to methanogens (which feed on lactate and acetate to produce methane), and **C** refers to methanotrophs (which consume CH_4 and O_2 to produce CO_2). The area shaded yellow in the figure, refers to the CO_2 produced by methanotrophs, by oxidizing CH_4 . The CO_2 and methane emissions are representative from previous literature (Corteselli2017).

to process the less usable ones. During this time the total CO_2 production falls, and due to the lack of sufficient oxygen, not all of the methane gets oxidized. This results in a net release of methane from the system.

Now as the process continues, the methane flux rises to a maximum, where the metabolites for methanogenesis are still available (from the accumulated overflow metabolism of the heterotrophs) but are getting depleted. After a certain tipping point, the methane flux also decreases, and the column becomes completely anoxic. In previous literature (Corteselli et al., 2017), such a rough pattern has been observed and the time frame for the whole process sums up to about one year. Seasonal rainfall mixes the anoxic waters with atmospheric oxygen well, and the cycle starts again by provision of pulsed high substrate input in the next fall. The process has been visualized in **Figure 2.5**.

Discussion

Here we considered the three major trade-offs of Warburg effect in a bacterial cell through linear equations and inequalities with the goal of improving the understanding of the ecological consequences of bacterial metabolism.

Comparison with previous work

Most previous modeling of the Warburg effect comes from the family of models called FBA (Flux Balance Analysis), which was briefly discussed in the introduction of this work (Molenaar et al., 2009; Moller et al., 2018; Schuster et al., 2015; Shlomi et al., 2011; Vazquez et al., 2010; Zhuang et al., 2011). These models use the knowledge of constructed biochemical networks, in particular the genome-scale metabolic network reconstructions (Molenaar et al., 2009). These networks consist of most identified metabolic reactions in a given organism and the

genes that encode each enzyme. The model estimates the flow of metabolites through this network, thus predicting the growth rate of an organism or rate of production of a given metabolite. These networks are specific to organisms and making a generalized model network for all aerobic heterotrophic bacteria would be an immense task. Calculations over such networks are also computationally intensive.

In contrast, we try to generate a very simple model using linear relations. The parameters that are used in the model can be inferred from experiments and are usable for any aerobic heterotrophic bacterial species. In addition, one can look at the growth and other estimates that are predicted by the model, and explore how they change as functions of the parameter values. Disadvantages of our approach are that we cannot predict the flux of other metabolites that are not encoded in our model and we cannot take protein costs and related terms into account.

Consequently, our model reflects the classic trade-off between generalizability and specificity. Because our aim was to study the ecology of these organisms, we chose a more general model that helps us explore the Warburg effect in this class of organisms broadly, rather than obtaining highly resolved pathways and fluxes. This simple formulation allowed us to incorporate the volume and surface area of the cell, leading to parameterization of shape and size of the cell. Exploration of how cellular metabolism and growth depend on cell shape and size (in addition to substrate concentration) is a novel feature of our model as compared to previous studies (Molenaar et al., 2009; Moller et al., 2018; Schuster et al., 2015; Shlomi et al., 2011; Vazquez et al., 2010).

Even though this family of FBA models has been noticeably successful in estimating fluxes, it is unable to explain overflow metabolism from first principles (Molenaar et al., 2009). Only when using measured fluxes like maximal oxygen uptake rates or other pathway capacities as auxiliary constraints do FBA models predict the use of fermentation or other inefficient routes

(Molenaar et al., 2009). Such a method does not seem adequate, as these constraints act like fitted parameters and the models do not fully explain why organisms simply do not increase the capacity of the efficient routes by producing more enzymes (Molenaar et al., 2009). We note, however, that Zhuang et al., (2011) incorporated membrane crowding and an upper limit on the number of respiratory enzymes in order to add a physical constraint on the model (FBAwmc).

In comparison, our model has a physical constraint similar to Zhuang et al., (2011), but unlike them, we use a generalized model instead of a specific organism's metabolic network. We take into consideration the facts that, as a component of the maintenance of cellular membrane structure in prokaryotes, the protein-to-lipid ratio is kept within certain bounds (Molenaar et al., 2009) and that at high rates of catabolism, the surface area available to proteins can become saturated and proteins (such as glucose transporters and OP enzymes) might 'compete' for expression (Zhuang et al., 2011).

Molenaar et al., (2009) used constraints of membrane structure maintenance and a maximal cellular density of proteins to construct their model. That work introduced a shape parameter but did not develop it further as the primary focus was to optimize enzyme systems in order to maximize growth. But even their model, which overcame some of the shortcomings of FBA, did not consider the minimal energy requirement of the cell.

Though minimal energy requirements for bacterial cells have been known for some time (Kempes et al., 2017), they were not directly incorporated in the models mentioned above. Our incorporation of such a constraint helped us to classify different metabolic regimes in bacteria, something that has not been done in many previous studies (Molenaar et al., 2009; Schuster et al., 2015; Shlomi et al., 2011; Vazquez et al., 2010). Vazquez et al., (2010) sought to identify dynamic regimes but ended up making two very strict ones: one with only OP and one with mixed

metabolism (fermentation and OP). Moller et al., (2018) aimed to classify dynamical regimes but did not take minimal energy requirements into consideration and did not address overflow metabolism.

Our adoption of a minimum energy constraint equation leads directly to the four regimes in metabolism, as described in the Results and **Figure 2.4**. These regimes stretch all the way from low substrate conditions to very high substrate concentrations. Even more, the same set of equations give us a direct estimate of overflow metabolism products. Understanding this set of regimes is crucial for many biotechnological applications which may utilize bacterial culture under different conditions.

Support from previous studies and experiments

There are numerous studies and experiments that follow the predictions outlined in our work - especially regarding the shape and size dependence of bacterial metabolism and growth.

In substrate poor conditions, it has been observed that *E. coli* reduces its Z by becoming much smaller to allow for more efficiency (Lange et al., 1991). In a similar note, certain strains of *E. coli* grew to become much larger than undiluted cells within a few generations (Begg et al., 1985) on serial dilution with fresh growth media.

A comparable result was shown in a long term evolutionary experiment in the same organism, where the mean cell volume increased in serial dilutions over 10,000 generations (Lenski et al., 1994). The increase in cell volume was also accompanied by an increase in relative fitness, which can be taken as a measure of r_3 , implying a lower Z has a higher r_3 , as predicted from our model. Such a general trend of higher fitness being associated with larger size was shown in about 40 species of obligate heterotrophic bacteria in another work (DeLong et al., 2010).

In a set of classic experiments, *Salmonella enterica* serovar Typhimurium produced wider cells (lower Z) in resource rich medium than when grown in minimal medium, and slow growing cells (low r_3) were smaller than fast growing cells (high r_3) (Schaechter et al., 1958). The latter was been also demonstrated in *E. coli*, where faster growing cells (high r_3) are significantly wider (high Z) (Nanninga et al., 1988).

Geobacter sulfurreducens has a capacity to process acetate more efficiently than a similar iron-reducer *Rhodospirillum rubrum* (Esteve-Nunez et al., 2005; Risso et al., 2009; Zhuang et al., 2011) due to its smaller size (and thus, larger Z). Similarly, living in energy-starved environments *Dehalococcus spp.* may have evolved its disc shape to maximize the dechlorination rate (given that the process has a low thermodynamic efficiency) (Jayachandran et al., 2004).

When grown in nutrient-limiting conditions, *Streptococcus* isolates grew as true filaments instead of as cocci (Ruoff, 1991; Frenkel and Hirsch, 1961). In a similar way, *Pseudomonas aeruginosa*, *Pseudomonas putida* and *Pseudomonas fluorescens* elongate into thin slim cells in nutrient poor environment, unlike the short rods observed in normal media (Shim et al., 1999; Steinberger et al., 2002).

A cell cannot increase its Z indefinitely by change in cell shape, hence another strategy would be to use subsidiary filaments to increase surface area (and Z). When deprived of certain substrates, *Actinomyces israelii* grows branched filamentous rods (high Z) and returns to its normal morphology once those substrates are added back (Pine and Boone, 1967). Nutrient-poor conditions also enhance filamentation in *Arthobacter globiformis* (Germida et al., 1980; Deutch and Perera, 1992) and *Clostridium welchii* (Webb, 1948, 1953}.

Further comments

As outlined in the Results, we can use a vectorized version of our model to understand the wetland gas emission dynamics via estimation of overflow metabolism. As a further speculation, we note that a similar process may have happened over longer geological timescales. In the past, events involving heavy releases of methane such as the end-Permian mass extinction (Rothman et al., 2014) might also be explainable to some extent using our formulation. High primary productivity and substrate availability due to increased CO_2 and weathering processes might have triggered overflow metabolism in the heterotrophic microbes leading to a large-scale production of metabolites such as acetate and lactate. These substrates can then be utilized by methanogens, especially acetate, which has been linked to the massive methanogenic burst at end-Permian (Rothman et al., 2014).

Although simple, our work provides a fairly straightforward way to look at ecological trends in systems of bacterial taxa. This is an added benefit which extends beyond the main results concerning overflow metabolism and the relative dynamics of respiration and fermentation. Not only are these trends supported by previous studies and experiments, but even more importantly they provide simple predictions, which can be tested in the lab.

Acknowledgements

WFF was partially supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant W911NF-14-1-0490.

References

1. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J.R. and Hwa, T., 2015. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature*, 528(7580), pp.99-104.
2. Bekker, M., De Vries, S., Ter Beek, A., Hellingwerf, K.J. and De Mattos, M.T., 2009. Respiration of *Escherichia coli* can be fully uncoupled via the nonelectrogenic terminal cytochrome bd-II oxidase. *Journal of bacteriology*, 191(17), pp.5510-5517.
3. Beg, Qasim K., Alexei Vazquez, Jason Ernst, Marcio A. de Menezes, Ziv Bar-Joseph, A-L. Barabási, and Zoltán N. Oltvai. "Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity." *Proceedings of the National Academy of Sciences* 104, no. 31 (2007): 12663-12668.
4. Begg, K.J. and Donachie, W.D., 1985. Cell shape and division in *Escherichia coli*: experiments with shape and division mutants. *Journal of bacteriology*, 163(2), pp.615-622.
5. Corteselli, E.M., Burtis, J.C., Heinz, A.K. and Yavitt, J.B., 2017. Leaf litter fuels methanogenesis throughout decomposition in a forested peatland. *Ecosystems*, 20(6), pp.1217-1232.
6. DeLong, J.P., Okie, J.G., Moses, M.E., Sibly, R.M. and Brown, J.H., 2010. Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life. *Proceedings of the National Academy of Sciences*, 107(29), pp.12941-12945.
7. De Deken, R.H., 1966. The Crabtree effect: a regulatory system in yeast. *Microbiology*, 44(2), pp.149-156.

8. Deutch, C.E. and Perera, G.S., 1992. Myceloid cell formation in *Arthrobacter globiformis* during osmotic stress. *Journal of applied bacteriology*, 72(6), pp.493-499.
9. Esteve-Núñez, A., Rothermich, M., Sharma, M. and Lovley, D., 2005. Growth of *Geobacter sulfurreducens* under nutrient-limiting conditions in continuous culture. *Environmental microbiology*, 7(5), pp.641-648.
10. Famili, I., Förster, J., Nielsen, J. and Palsson, B.O., 2003. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences*, 100(23), pp.13134-13139.
11. Flamholz, A., Noor, E., Bar-Even, A., Liebermeister, W. and Milo, R., 2013. Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proceedings of the National Academy of Sciences*, 110(24), pp.10039-10044.
12. Frenkel, A. and Hirsch, W., 1961. Spontaneous development of L forms of streptococci requiring secretions of other bacteria or sulphhydryl compounds for normal growth. *Nature*, 191(4789), pp.728-730.
13. Frick, T. and Schuster, S., 2003. An example of the prisoner's dilemma in biochemistry. *Naturwissenschaften*, 90(7), pp.327-331.
14. Germida, J.J. and Casida Jr, L.E., 1980. Myceloid growth of *Arthrobacter globiformis* and other *Arthrobacter* species. *Journal of bacteriology*, 144(3), pp.1152-1158.
15. Jayachandran, G., Görisch, H. and Adrian, L., 2004. Studies on hydrogenase activity and chlorobenzene respiration in *Dehalococcoides* sp. strain CBDB1. *Archives of microbiology*, 182(6), pp.498-504.
16. Kareva, I., 2011. Prisoner's dilemma in cancer metabolism. *PloS one*, 6(12), p.e28576.

17. Kempes, C.P., van Bodegom, P.M., Wolpert, D., Libby, E., Amend, J. and Hoehler, T., 2017. Drivers of bacterial maintenance and minimal energy requirements. *Frontiers in microbiology*, 8, p.31.
18. Lange, R. and Hengge-Aronis, R., 1991. Growth phase-regulated expression of *bolA* and morphology of stationary-phase *Escherichia coli* cells are controlled by the novel sigma factor sigma S. *Journal of Bacteriology*, 173(14), pp.4474-4481.
19. Lenski, R.E. and Travisano, M., 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences*, 91(15), pp.6808-6814.
20. Lunt, S.Y. and Vander Heiden, M.G., 2011. Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annual review of cell and developmental biology*, 27, pp.441-464.
21. McInerney, M.J. and Bryant, M.P., 1981. Anaerobic degradation of lactate by syntrophic associations of *Methanosarcina barkeri* and *Desulfovibrio* species and effect of H₂ on acetate degradation. *Applied and environmental microbiology*, 41(2), pp.346-354.
22. Molenaar, D., Van Berlo, R., De Ridder, D. and Teusink, B., 2009. Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular systems biology*, 5(1), p.323.
23. Möller, P., Liu, X., Schuster, S. and Boley, D., 2018. Linear programming model can explain respiration of fermentation products. *PloS one*, 13(2), p.e0191803.
24. Nanninga, N., 1988. Growth and form in microorganisms: morphogenesis of *Escherichia coli*. *Canadian journal of microbiology*, 34(4), pp.381-389.

25. Newsholme, E.A., Crabtree, B. and Ardawi, M.S.M., 1985. The role of high rates of glycolysis and glutamine utilization in rapidly dividing cells. *Bioscience reports*, 5(5), pp.393-400.
26. Pfeiffer, T., Schuster, S. and Bonhoeffer, S., 2001. Cooperation and competition in the evolution of ATP-producing pathways. *Science*, 292(5516), pp.504-507.
27. Pine, L. and Boone, C.J., 1967. Comparative cell wall analyses of morphological forms within the genus *Actinomyces*. *Journal of bacteriology*, 94(4), pp.875-883.
28. Price, N.D., Reed, J.L. and Palsson, B.Ø., 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11), pp.886-897.
29. Risso, C., Sun, J., Zhuang, K., Mahadevan, R., DeBoy, R., Ismail, W., Shrivastava, S., Huot, H., Kothari, S., Daugherty, S. and Bui, O., 2009. Genome-scale comparison and constraint-based metabolic reconstruction of the facultative anaerobic Fe (III)-reducer *Rhodoferrax ferrireducens*. *BMC genomics*, 10(1), pp.1-19.
30. Rothman, D.H., Fournier, G.P., French, K.L., Alm, E.J., Boyle, E.A., Cao, C. and Summons, R.E., 2014. Methanogenic burst in the end-Permian carbon cycle. *Proceedings of the National Academy of Sciences*, 111(15), pp.5462-5467.
31. Ruoff, K.L., 1991. Nutritionally variant streptococci. *Clinical microbiology reviews*, 4(2), pp.184-190.
32. Ruppin, E., Papin, J.A., De Figueiredo, L.F. and Schuster, S., 2010. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Current opinion in biotechnology*, 21(4), pp.502-510.

33. Schaechter, M., Maaløe, O. and Kjeldgaard, N.O., 1958. Dependency on medium and temperature of cell size and chemical composition during balanced growth of *Salmonella typhimurium*. *Microbiology*, 19(3), pp.592-606.
34. Schuster, S., Boley, D., Möller, P., Stark, H. and Kaleta, C., 2015. Mathematical models for explaining the Warburg effect: a review focussed on ATP and biomass production. *Biochemical Society Transactions*, 43(6), pp.1187-1194.
35. Shehata, T.E. and Marr, A.G., 1971. Effect of nutrient concentration on the growth of *Escherichia coli*. *Journal of bacteriology*, 107(1), pp.210-216.
36. Shim, H. and Yang, S.T., 1999. Biodegradation of benzene, toluene, ethylbenzene, and o-xylene by a coculture of *Pseudomonas putida* and *Pseudomonas fluorescens* immobilized in a fibrous-bed bioreactor. *Journal of Biotechnology*, 67(2-3), pp.99-112.
37. Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R. and Ruppin, E., 2011. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS computational biology*, 7(3), p.e1002018.
38. Steinberger, R.E., Allen, A.R., Hansma, H.G. and Holden, P.M., 2002. Elongation correlates with nutrient deprivation in *Pseudomonas aeruginosa* unsaturated biofilms. *Microbial ecology*, 43(4), pp.416-423.
39. Swain, A. and Chatterjee, S., 2017. A new formulation for determination of the competition coefficient in multispecies interaction for Lotka—Volterra type competition models. *Current Science*, pp.1920-1926.
40. Tseng, Ching-Ping, Jeffery Albrecht, and Robert P. Gunsalus. "Effect of microaerophilic cell growth conditions on expression of the aerobic (cyoABCDE and cydAB) and

anaerobic (narGHJI, frdABCD, and dmsABC) respiratory pathway genes in *Escherichia coli*." *Journal of bacteriology* 178, no. 4 (1996): 1094-1098.

41. Vander Heiden, M.G., Cantley, L.C. and Thompson, C.B., 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930), pp.1029-1033.
42. Vazquez, A., Liu, J., Zhou, Y. and Oltvai, Z.N., 2010. Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited. *BMC systems biology*, 4(1), pp.1-9.
43. Vemuri, G.N., Altman, E., Sangurdekar, D.P., Khodursky, A.B. and Eiteman, M.A., 2006. Overflow metabolism in *Escherichia coli* during steady-state growth: transcriptional regulation and effect of the redox ratio. *Applied and environmental microbiology*, 72(5), pp.3653-3661.
44. Voet, D. and Voet, J.G., 2004. *Biochemistry*. 3rd edition, Wiley and Sons, Hoboken.
45. Warburg, O., 1956. On the origin of cancer cells. *Science*, 123(3191), pp.309-314.
46. Webb, M., 1948. The Influence of Magnesium on Cell Division. 1. The Growth of *Clostridium welchii* in Complex Media Deficient in Magnesium. *Journal of General Microbiology*, 2(3), pp.275-87.
47. Webb, M., 1953. Effects of magnesium on cellular division in bacteria. *Science*, 118(3073), pp.607-611.
48. Young, K.D., 2006. The selective value of bacterial shape. *Microbiology and molecular biology reviews*, 70(3), pp.660-703.
49. Zhuang, K., Vemuri, G.N. and Mahadevan, R., 2011. Economics of membrane occupancy and respiration. *Molecular systems biology*, 7(1), p.500.

Chapter 3 : Were Neoproterozoic atmospheric methane hazes and early Paleoproterozoic glaciations driven by the rise of oxygen in surface environments?

Abstract

Geochemical evidence suggests methane was the predominant greenhouse gas in the Archean and early Proterozoic eons. Consequently, fluctuations in methane concentration in Earth's atmosphere (i.e., 'methane hazes') would have contributed to climate change and influenced the flux of UV radiation reaching surface environments. If correct, understanding what factors (e.g., O_2 and/or resource concentration) drove the biological methane cycle might shed light on the repetition of biological, atmospheric and climatic events preserved in the sedimentary rock record from ~2.8 to 2.0 BYA. To explore these interdependencies, we developed a novel dynamical model of microbial ecological interactions to investigate the conditions under which methane is preferentially released to the atmosphere. We found that the interplay between resource and O_2 availability results in complex cyclic methane dynamics unrelated to the functional groups or efficiencies of microbial communities, to initial conditions, or to other model constraints. Based on these results, we propose that the cyclicity of methane haze events and glacial episodes in the Neoproterozoic and early Paleoproterozoic may have been linked to the progressive increase in oceanic and atmospheric O_2 through the interval.

Introduction

Like Titan, Saturn's orange-shrouded moon, the early Earth may have been episodically enveloped in a thick blanket of atmospheric methane (CH_4), a powerful greenhouse gas produced by archaeal methanogens in anoxic environments. A methane haze would have formed if the CH_4/CO_2 ratio was >0.2 (Trainer et al., 2004; Domagal-Goldman et al., 2008) contrasting with the value of 0.005 in the modern oxygenated atmosphere. Although the depth and distribution of the CH_4 haze during such events likely varied across the planet due to vertical atmospheric motion and mixing (Wolf and Toon, 2010), the biogenic gas would nonetheless have warmed the troposphere, and through attenuation of the solar UV flux, protected surface environments from high energy radiation.

Carbon isotope evidence from Neoproterozoic (2.8 to 2.5 BYA) organic matter preserved in sedimentary rocks (Eigenbrode and Freeman, 2006) suggests a prolonged increase in the production of CH_4 and its consumption by methanotrophs to form biomass. Depending on the leak rate of the greenhouse gas from anoxic sediments, CH_4 may have accumulated in sufficient concentrations in the atmosphere to form periodic hazes. The extent and depth of such hazes is likely to have been modulated by the photosynthetic production and accumulation of molecular oxygen (O_2) in the oceans and atmosphere.

Due to the rapid oxidation of CH_4 to CO_2 in the presence of O_2 (Saunio et al., 2016), the atmospheric concentration of CH_4 is generally thought to represent a balance between source and sink reactions. Tipping the scales between the two would thus modulate the size of the atmospheric reservoir of CH_4 . On the source side, the production of CH_4 could have been affected by competition for critical metabolic substrates between methanogens and sulfate reducers, or by the lack of key metal cofactors, like nickel, required for methanogenic enzymes (Anbar and Knoll, 2002; Konhauser et al., 2009). In this analysis, however, we consider ecological factors and biological interactions that may have increased the CH_4 flux from surface environments that are

directly associated with the release of photosynthetic O₂. Such a scenario would be particularly relevant in the Neoproterozoic when periodic CH₄ hazes are suggested from geochemical measurements of sedimentary rocks and photochemical models (Zerkle et al., 2012; Farquhar et al., 2013; Ikon et al., 2015; 2017, Williford et al., 2016).

Our dynamic resource-driven model for CH₄ generation and release is based upon the metabolic repercussions of the Warburg Effect in microbial functional groups (see Swain and Fagan, 2018) – which was initially proposed to explain uncontrolled cellular proliferation during cancer growth (Warburg, 1956) – to an ecological scenario appropriate to Neoproterozoic and early Paleoproterozoic surface environments and atmospheric fluctuations. The Warburg Effect in microorganisms describes a shift in metabolic respiration where the energetic process of oxidative phosphorylation (OP), which produces CO₂ as an end product, is replaced by glycolysis that instead generates products that can be incorporated into biomass but yields 10 times less energy. This shift in respiration mechanisms is particularly pronounced during the rapid growth of microbial cells that are not resource limited – even in the presence of oxygen. This metabolic phenomenon has been observed in a wide variety of organisms such as bacteria (where it is termed “overflow metabolism”; Basan et al., 2015, Schuster et al., 2015; Zhuang et al., 2011), yeast (the “Crabtree effect”; Deken, 1966), and a diversity of human cells, including Kupffer and microglial cells (specialized macrophages in the liver and brain, respectively), lymphocytes (white blood cells), and fast twitch muscle fibers (Schuster et al., 2015; Vander Heiden et al., 2009). Especially in microorganisms, this phenomenon is known to have cascading effects on the stoichiometry and community dynamics in natural ecosystems (see Manzoni et al., 2012; Carlson et al., 2018; Kreft et al., 2020). While the Warburg Effect may seem energetically counterintuitive, one can observe that as cells grow rapidly, the kinetic advantage of glycolysis that occurs in the cytoplasm results

in a two-fold energy advantage over OP, which is restricted to the cell membrane (Voet and Voet, 2004). Past models thus describe a metabolic arms race between OP and glycolysis that includes: i) efficiency vs. rate (OP is more efficient, but glycolysis is faster), ii) surface area vs. volume (glycolysis occurs in the cell interior whereas OP is constrained to the cell surface), and iii) biomass production (glycolysis produces organic substrates that can be used in biomass production directly, which is generally not true for OP) (see Swain and Fagan, 2018; Schuster et al., 2015). Results from these models suggest that large inputs of resources (i.e., any carbon substrate that allows the microbes to grow) with even small inputs of oxygen can stimulate glycolysis over the more energetic process of OP, and therefore, can affect ecological processes, specifically methanogenesis, that use these glycolytic overflow products (e.g., formate, lactate, and acetate).

To understand the dynamics of how resource and O₂ availability can affect the production and consumption of CH₄, we present a dynamical model of microbial ecological interactions, using the concept of overflow metabolism associated with the Warburg Effect. The ecological interactions described in this ordinary differential equation (ODE) model may be relevant to geological episodes where pulsed resource inputs or the availability of O₂ can alter the CH₄ production rate. Our ODE model might then apply to climatic and environmental perturbations of the early Earth, including the cyclic patterns of Neoproterozoic CH₄ haze events (Pavlov et al., 2001; Zerkle et al., 2012; Izon et al., 2015; 2017) and early Paleoproterozoic glacial episodes (e.g., Bekker and Kaufman, 2007).

In our model, we focus on five major functional groups of microbes: obligate aerobic heterotrophic bacteria (A), methanogens (B), aerobic methanotrophs (C), generalized fermenters (F), anaerobic methanotrophs (G; also known as the anaerobic oxidation of methane (AOM) consortia), and anaerobic heterotrophs (H; such as iron and sulfur reducers). All of these

microbial functional groups are involved in the production and/or consumption of generalized resources (R), glycolytic/fermentation products (FP), O₂, CH₄, and CO₂ (**Fig. 3.1**). Here, the generalized resources are an abstract representation of a variety of materials that organisms might use as a carbon source, or in an indirect assimilation of carbon source and growth. Through this generalization of resources, we allow that some carbon-based compounds come from oxygenic photosynthesis. For example, anoxygenic photosynthesis (which, in turn, depended on and was limited by electron acceptors; see Ward et al., 2019) appears to have provided carbon sources for microbes, especially in Neoproterozoic and early Paleoproterozoic sedimentary environments. As a consequence of this generalization of resources, our model considers a decoupled system of R and O₂. In addition, non-carbon resources can vary in their stoichiometry in organisms (see Sterner and Elser, 2002); therefore, this decoupling of R and O₂ also helps to decode the effects of the continental nutrient influx to the oceans, or related phenomena.

Key relationships within the flow diagram (**Fig. 3.1**) can be understood as follows. Methanogens depend upon fermentation products (FP) for their methane production; FP are simple organic substrates such as formate, lactate, and acetate that are produced by Fermenters (F). Aerobic heterotrophic bacteria (A) also produce fermentation products (FP) through overflow metabolism, insofar as this functional group responds to increasing R by producing more FP than CO₂, and thereby stimulating methanogenesis (see **Methods**). Anaerobic heterotrophs (H) compete with methanogens (B) for FP. Methane is consumed by aerobic methanotrophs (C) that utilize O₂ to produce CO₂, and by anaerobic methanotrophs (G), in consortia with various other groups of microbes, to produce either alkalinity or CO₂ through the anaerobic oxidation of methane

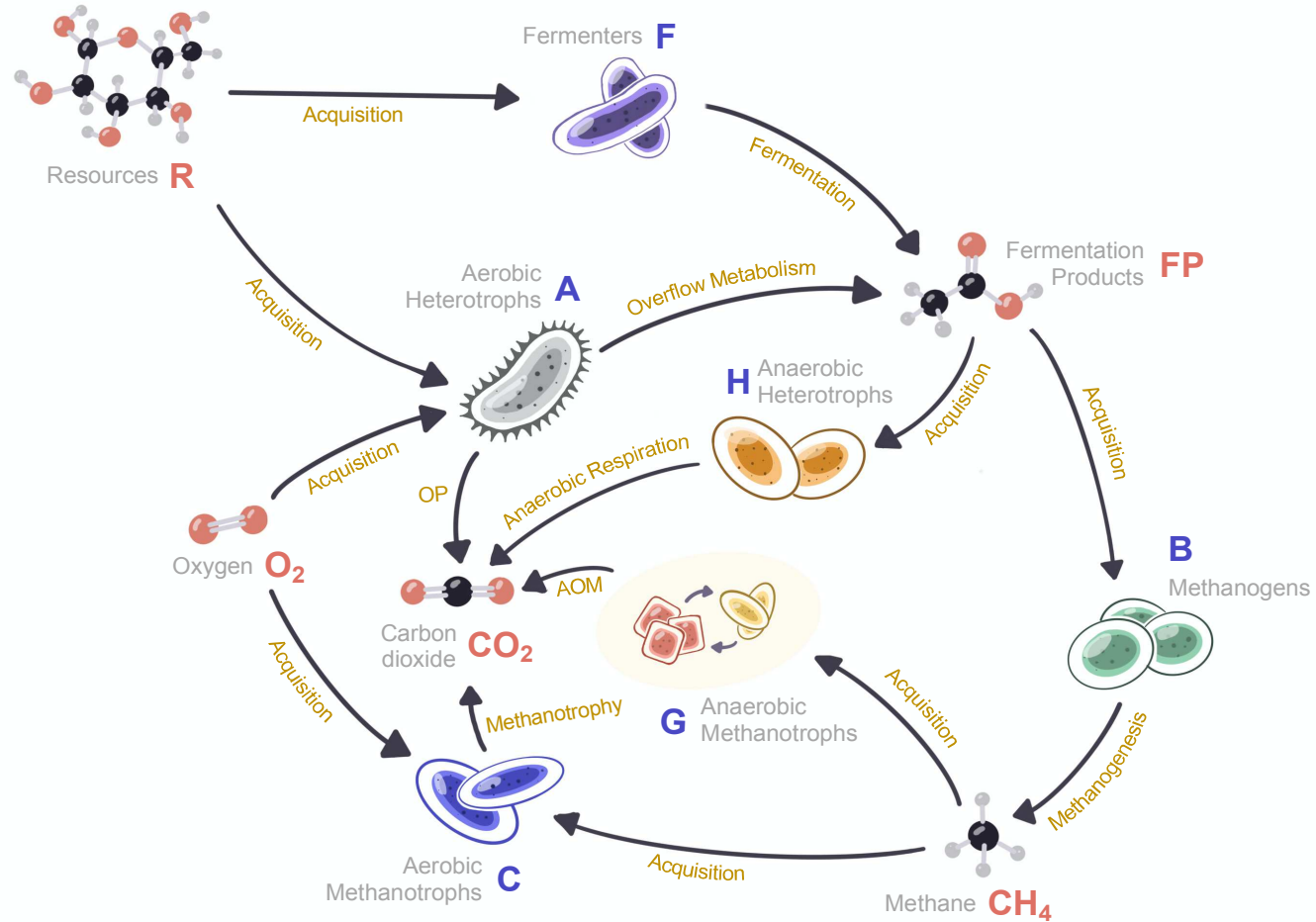


Figure 3.1. Flow diagram representation of the model system. Here (in blue) A, B, C, F, G and H are functional groups of heterotrophic bacteria, methanogens, aerobic methanotrophs, fermenters, anaerobic methanotrophs, and anaerobic heterotrophs, respectively. The flux arrows indicate processes that produce (in red) either metabolic reactants or gases cycled among the microbial functional groups. Ordinary differential equations with detailed explanations representing this dynamical system appear in **Methods**.

(AOM). Changes in the size of microbial functional groups are modelled as functions of the availability of R and the efficiency of their utilization (energy production), as well as their basal energy requirements necessary to maintain homeostasis (energy requirement). We assume methanogens (B) represent both acetoclastic and hydrogenotrophic methanogenesis with respect to reduced FP compounds. Acetoclastic methanogens use primary FP, and hydrogenotrophic methanogens depend on secondary fermentation for H_2 and formate. We did not include the hydrogenotrophic methanogens as a sink for CO_2 , however, because this can cause unintended model estimation errors that are important for the decoupling between R and O_2 . As we are assuming a decoupling between CO_2 , O_2 , and R (given the unknown extent of oxygenic photosynthesis), we can either leave CO_2 as a sink (as modeled) or couple CO_2 back to R as a function of oxygenic and anoxygenic autotrophy. In the latter case, we would include an explicit CO_2 flux to the hydrogenotrophic methanogens. In the former (as chosen), we assume only that there is enough CO_2 to fulfill all necessary demands, and therefore the substrate is not a limiting reagent.

The main rationale behind constructing this coupled ODE-based ecological model is to understand the response of microbial functional groups to the availability of R and O_2 , through mutual interactions and usage of substrates. Changes in the abundance of resources, metabolic products (FP, CH_4 , and CO_2), and O_2 associated with consortia interactions are functions of efficiencies of i) overflow metabolism, ii) respiration, and iii) biomass production and maintenance, as well as the rate of addition of both resources and O_2 . To explore the model, we considered both driven and non-driven states. First, for non-driven dynamics, we explore the basic dependence of metabolic parameters and species interactions in closed systems wherein microbes and resources start from an initial condition, and are then closed off to external changes. Second, we consider driven

dynamics in open systems, in which there are constant rates of addition of R and O₂. Numerical exploration of the model suggests that the interplay between the two results in complex, but robust, cyclic methane dynamics. We relate this new dynamic perspective to the episodic nature of Neoproterozoic methane haze events and early Paleoproterozoic glaciation.

Results

No External Forcing

For the case of no external forcing (non-driven) where the only initial inputs were R and O₂, we ran 500,000 model simulations to steady-state (**Figure S3.1**). Across these simulations, we varied key model parameters (the efficiency of various metabolic reactions, including methanogenesis, AOM, and fermentation) and initial conditions (starting populations of different microbial functional groups, and original abundance of R and O₂) using Latin Hypercube Sampling (LHS; Carnell, 2020) to explore their impacts on final gas (O₂, CO₂, and CH₄) concentrations. We gauged the importance of the different parameters and conditions using partial rank correlation coefficients (PRCC) (see **Methods**), and found that steady-state CH₄ and CO₂ concentrations varied the most based upon the metabolic parameters of methanogens (these included resource incorporation efficiency (β), resource utilization efficiency (κ) and energy requirement (u)) (**Figure S3.2B,C**). The model further reveals that steady state CO₂ concentration varied strongly as a function of the metabolic parameters that controlled anaerobic methanotrophy (or AOM, G) (**Figure S3.2B**). Steady state O₂ concentrations varied as functions of the metabolic parameters of aerobic methanotrophs (C), aerobic heterotrophic bacteria (A), and methanogens (B) (**Figure S3.2A**). None of the initial conditions that we varied (i.e., starting populations of different microbial functional groups and initial concentrations of R and O₂) had a high PRCC for any of the stable

state gas concentrations (**Figure S3.3**), except for the dependence of the final CO₂ concentration on the initial value of O₂ concentration (**Figure S3.3B**).

Although the steady state gas concentrations varied quantitatively with certain parameters, the qualitative trends that they depicted were relatively invariant across the Latin Hypercube sweep of parameter space and initial conditions. **Figure 3.2** depicts a specific case, where we plot the steady state variations of the proportions of various gases in a closed, non-driven system (initialized with a biologically-relevant base parametrization) due to changes in the methane production efficiency coefficient (β , see Eqs. 4, 9, and 11 in **Methods**), which was set at 1%, 10%, and 50%. With increasing efficiency, the system became more anoxic (lower O₂) and less acidic (lower CO₂), while CH₄ concentrations increased, especially at low initial O₂ and high initial R inputs. We recorded similar model responses for variations of all other parameters (See **Figures S3.4A-O**). However, we noted lower than expected CO₂ concentrations at low initial O₂ inputs for low efficiencies of parameters governing the incorporation and utilization of resources by the aerobic heterotrophs (see Eqs. 2, 3, 8, and 10 in **Methods**; and **Figs. S3.4D and S3.4H**); these can be attributed to the PRCC dependence of CO₂ on initial O₂ concentration. We also noted a minor deviation in CH₄ concentration at a high initial O₂ input in the Ω (resource utilization efficiency for anaerobic methanotrophs, G; see **Methods**) simulations (see Eqs. 7 and 10 in **Methods**; **Fig. S3.4K**); this was attributable to the formation of CO₂ and its dependence on metabolic parameters of G. These non-linearities arose in cases where the efficiency of usage, or of incorporation, is either biologically unrealistic, or are at the extreme end of what is possible for biological systems (e.g., a very low O₂ affinity for aerobic heterotrophs or a high CH₄ usage efficiency in anaerobic methanotrophs (see Morris and Schmidt 2013; Bowles et al., 2019). Overall, these results signify

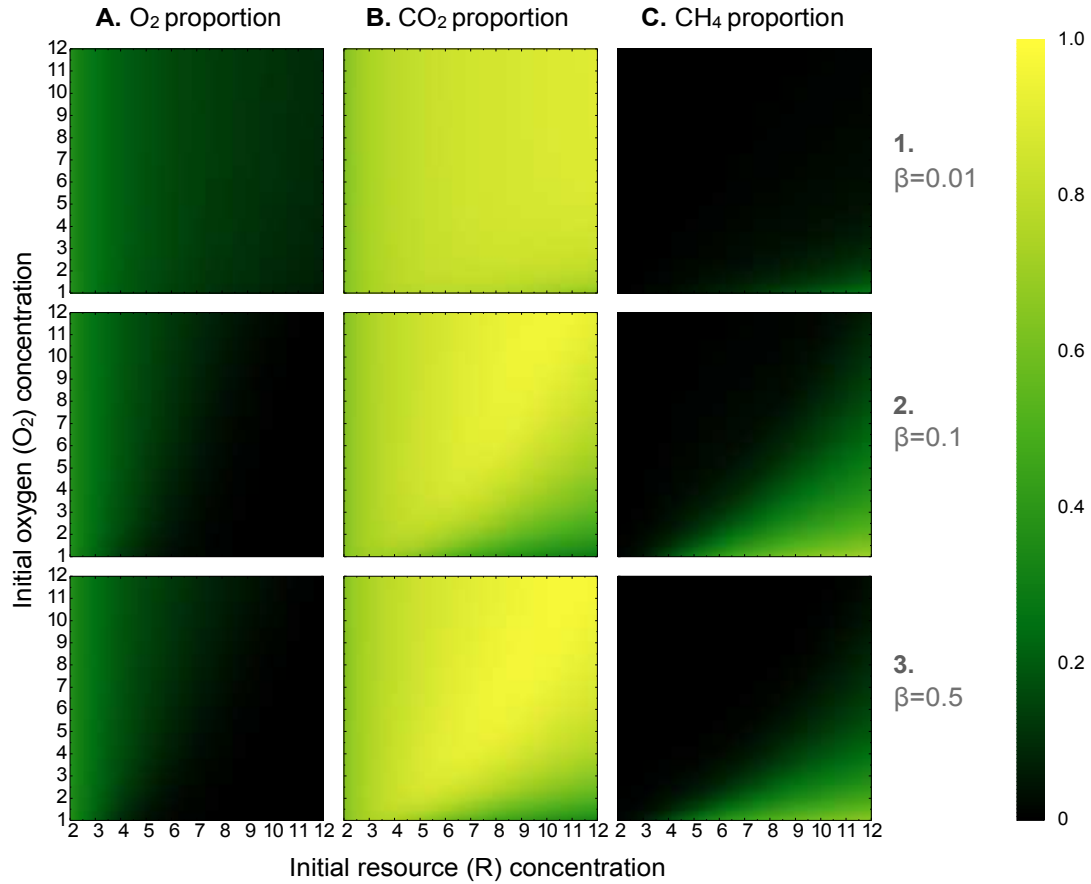


Figure 3.2 Dynamic variations in O₂, CO₂, and CH₄ proportions (in columns A-C) run to steady state at set O₂ and R input values, with the coefficient of CH₄ production (β) set at 1, 10, and 50% in rows 1-3, respectively. Note that the resource (R) axis begins at 2; this is due to constraints posed by the minimal energy requirements for each microbial functional group, below which all functional groups diminish.

that, on average, the qualitative outputs of our ODE computations remained consistent across the parametrizations and initial conditions in the non-driven setting. We further tested the stability and the sensitivity of the modeled system to the addition of new microbial functional groups to the main structure by comparing with results of a model lacking H (anaerobic heterotrophs) that compete directly with B (methanogens). The comparison shows no significant differences among the two models in terms of their partial rank correlation coefficients (see **Figure S3.2**).

External Forcing

In contrast with the no external forcing scenario above, microbial systems may also be forced, particularly by influxes of metabolizable organic substrates (resource, R) and O₂ that are linked through oxygenic photosynthesis, wherein R is synthesized and O₂ is released as a byproduct of the metabolic activity. Model runs with only forced R inputs were qualitatively simple. As illustrated in **Fig. S3.5**, initiating the model with a constant rate of addition of R depletes O₂ rapidly regardless of its initial abundance. Under these conditions, the Warburg Effect in oxidative heterotrophs does not play a major role in the long-term dynamics of the microbial system. In this scenario, the Warburg Effect only affects the system initially, so that most FP is derived from general fermenters (F) for the subsequent production of CH₄ through methanogenesis. Most notably, simulations with lower initial O₂ (**Fig. S3.5B**) resulted in an oscillatory behavior over longer periods of CH₄ generation. These results contrast with simulations with higher initial O₂ (**Fig. S3.5A**), which resulted in rapid damping of the oscillatory behavior. Note, however, that all simulations with the same parameter values reached the same steady state irrespective of the initial conditions, as one can extrapolate from **Figure S3.5** (and PRCC values from steady state dynamics). Thus, these simulations revealed that differences in the initial O₂ input only affected the relative timescale of events, as the saturation values (amplitudes) and

profiles of the CH₄, CO₂ and R concentrations remained the same. Higher initial O₂ stabilized the system more rapidly to steady state condition, while higher initial R under low O₂ resulted in sustained oscillations over longer intervals. If the initial R input is held constant, the dynamic solution to forcing the microbial system with a fixed rate of addition of O₂ is that CH₄ will be produced briefly associated with the initial concentration of resource, and O₂ will fall as long as methane is produced and aerobic methanotrophs consume it (**Fig. S3.6**). After the CH₄ resource is depleted, the continued input of O₂ overwhelms the system and kills the anaerobes, thereby allowing O₂ to increase at a linear rate. Different starting concentrations of R have no apparent effect on system-level dynamics.

Truly complex dynamics occur when both R and O₂ are supplied at constant rates (the input rate per unit time, denoted as $\sigma_1(t)$ for R and $\sigma_2(t)$ for O₂; see Eqs. 8 and 12 in **Methods**) (**Fig. 3.3**). Low oxygen supply rates relative to resource supply rates (i.e., high $\sigma_2(t) / \sigma_1(t)$) or conversely low $\sigma_2(t) / \sigma_1(t)$) result in large amplitude, slow frequency oscillations in methane and oxygen concentrations in the system, a phenomenon that differs from the model behavior with only resource addition (see **Figure S3.5; Figure 3.3A**). With decreasing $\sigma_2(t) / \sigma_1(t)$, these oscillations reduce in amplitude and increase in frequency until a threshold value (of relative O₂ input), where the system reaches a non-oscillatory stable state (**Figure 3.3A-J; Figure 3.4**). The transitions into and out of the periodic behavior appear smooth, and the amplitude and periodicity of the CH₄ peaks within the dynamic interval decline with increasing O₂ input rate. Moreover, the ratio of resource to oxygen input rates at which oscillatory behavior transitions to a stable, non-oscillatory pattern remains constant across the parameter variations and initial conditions (i.e. the ratio of R to O₂ input remained at 5, or, $\sigma_2(t) / \sigma_1(t) \sim 5$) (**Figure 3.4B**). **Figure 3.4A** further illustrates the relationship between the O₂ input rate (σ_2 , varied from 10 to 20% of σ_1) on the

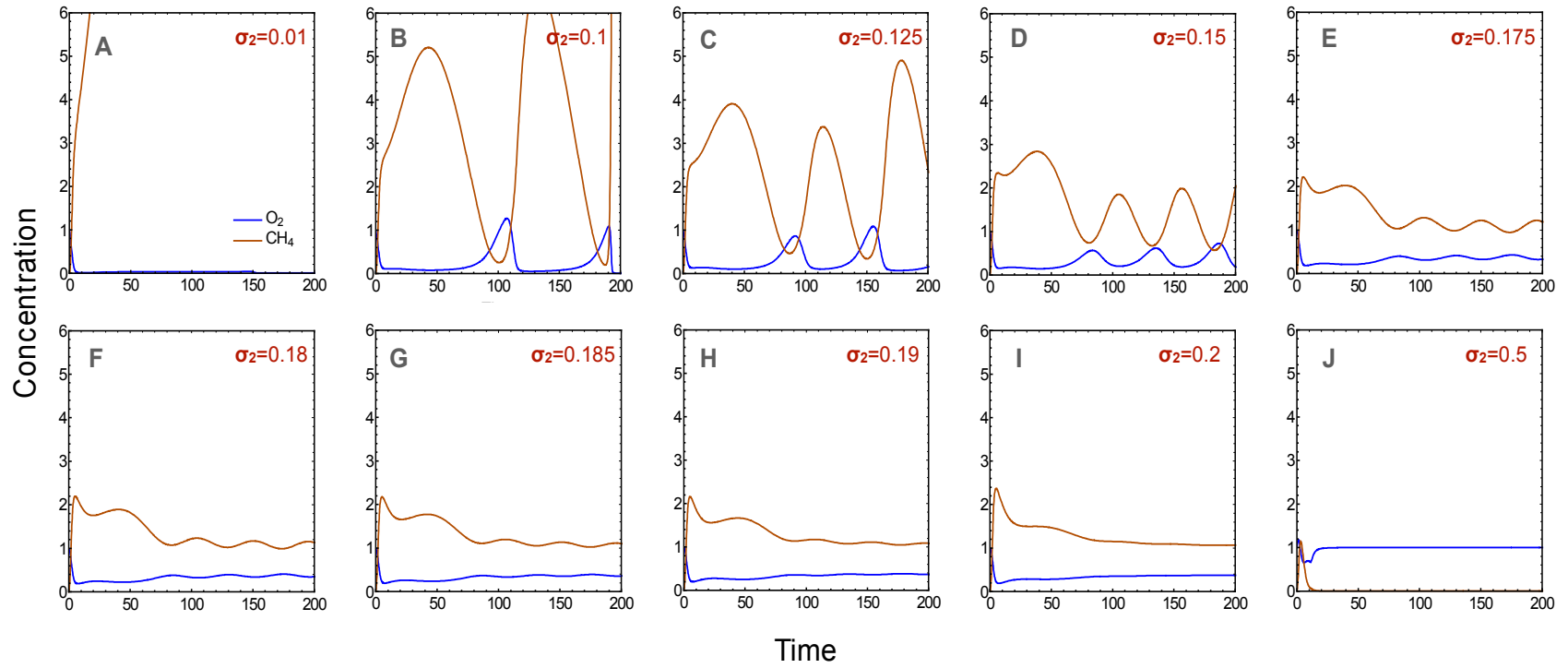


Figure 3.3: Dynamic variations in O_2 and CH_4 for the externally forced model with constant rates of addition of R and O_2 . The O_2 input rate (σ_2) varied between 1% to 50% relative to the resource input rate across panels (σ_1), but in all cases the R input rate (σ_1) was held constant at 1.

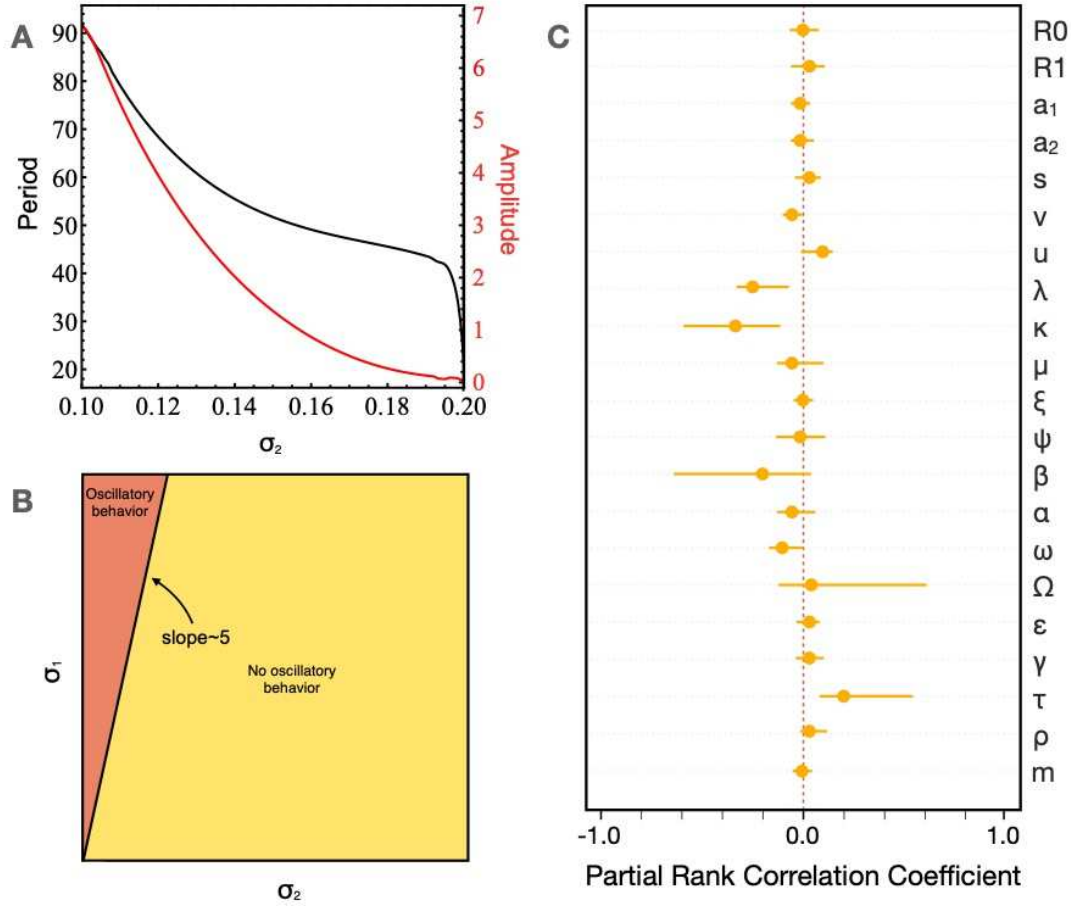


Figure 3.4: (A) shows the dependence of amplitude (in red) and time period (in black) of the CH₄ oscillations in Fig. 3.3 as a function of the O₂ input rate (σ_2). (B) demarcates the resource input (σ_1) -oxygen input (σ_2) space on the basis of oscillatory behavior for our standard set of parameters. (C) depicts the dependence of the slope of the line separating the space between oscillatory and non-oscillatory behavior in σ_1 - σ_2 space on specific parameters (for description of parameters, see **Methods**; **Table 3.1** in **Methods**) over a wide range of parameter combinations through partial rank correlation coefficient (PRCC).

amplitude and periodicity of CH₄ peaks. Clearly the greatest amplitudes and periods in CH₄ concentration occur at low O₂ input rates, and these both decline rapidly as oxygen overwhelms the system, leading to a phase transition when σ_2 is $\sim 20\%$ of σ_1 (**Figures 3.3 and 3.4**).

We conduct a similar robustness test for the dynamics of a reduced model (lacking the H functional group of anaerobic heterotrophs) with the dynamics of the full model, and find negligible differences between the two (e.g., **Figure S3.7**). This signifies that the model is mostly affected by the relative addition rates of oxygen to resources, and not by other parameters or initial conditions (see **Figure S3.2, S3.3, S3.7**) for large qualitative changes.

We also perform a sensitivity analysis of the slope demarcating the oscillatory and non-oscillatory behavior in a variety of parameter combination (selected through LHS) with the results illustrated in **Figure 3.4C**. We find that the space demarcation is affected most by the resource utilization efficiency of methanogens (β), anaerobic methanotrophs (Ω) and anaerobic heterotrophs (λ), as well as resource incorporation efficiency of aerobic methanotrophs (λ) and methanogens (κ). Increased conversion efficiency of carbon substrates to methane (represented by β) has been hypothesized to result in severe biotic crises and methanogenic bursts in the past (Rothman et al., 2014). The role of aerobic and anaerobic methanotrophs (represented by Ω and λ , *respectively*) is usually to control the large-scale methane flux, and that of anaerobic heterotrophs is to compete for carbon substrates with methanogens (Lyu et al., 2018).

Discussion

Our dynamic, resource-driven ecological model based on the Warburg Effect reveals clues to the production of CH₄ in response to the availability of molecular O₂ when applied to a consortium of aerobic and anaerobic microbial functional groups. By progressively increasing the input rate of O₂ from 10-20% with respect to the resource input rate, the overflow metabolism of the

heterotrophic bacteria provides the resources necessary to produce CH₄ in greater proportions than it is consumed by both aerobic and anaerobic methanotrophs.

Applied to potentially coupled biological and environmental events across the Neoproterozoic-early Paleoproterozoic transition, this dynamic ecological model may help to explain how the evolution of oxygenic photosynthesis drove fluctuations in the oxidation state of surface environments, as well as the pulsed releases of CH₄ to the atmosphere that formed periodic methane hazes. Carbon isotope compositions of organic matter preserved in sedimentary rocks deposited between ~2.8 to 2.5 billion years ago reveal unprecedented levels of ¹²C enrichment that are most parsimoniously attributed to the incorporation of biomass formed by methanotrophs (Hayes, 1994; Eigenbrode and Freeman, 2006). Under aerobic conditions, CH₄ is oxidized with molecular O₂ to form biomass through the serine or ribulose monophosphate pathways, and CO₂ (Chistoserdova et al., 2005), where the minimum concentrations of CH₄ and O₂ required to drive the microbial process are estimated at 5-10 μM and 12-20 μM, respectively (Harris and Hanson, 1980; Lidstrom and Somers, 1984). Under anaerobic conditions NO₃⁻, NO₂⁻, SO₄²⁻, Fe³⁺, or Mn²⁺ may also be used to oxidize CH₄ to form simple organic compounds, like acetate (Yan et al., 2018). If photosynthetically derived resources were high — as expected by the pervasive occurrence of thick organic-rich black shales and carbonates in Neoproterozoic sedimentary successions worldwide (e.g. Zhelezinskaia et al., 2014) — then almost any local increase in O₂ could have initiated overflow metabolism in extant heterotrophic bacteria. Through microbial community dynamics, this could have resulted in the enhanced release of CH₄ to the atmosphere, as described in the model above. Distributed globally, oceanic O₂ oases could then have been the source of Neoproterozoic CH₄ hazes.

Independent evidence for the presence of repetitious Neoproterozoic CH₄ haze events comes from multiple sulfur isotope measurements of pyrite from the same samples where organic matter was isolated and analyzed. In Neoproterozoic marginal marine sedimentary successions on multiple continents, carbon isotope evidence links enhanced methanotrophic biomass inputs in organic-rich sedimentary rocks to the presence of a CH₄ haze that would have attenuated the UV flux and isotopically altered the products of atmospheric SO₂ photochemistry (see Domagal-Goldman et al., 2008), predominantly sulfate aerosols and elemental sulfur (Kasting, 2001). The photochemical reactions resulted in mass-independent fractionation, or redistribution, of the sulfur isotope masses (S-MIF) in both of the products that rained down and accumulated in surface environments. Notable deviations in the ratio of multiple rare sulfur isotope abundances ($\Delta^{33}\text{S}/\delta^{34}\text{S}$ and $\Delta^{36}\text{S}/\Delta^{33}\text{S}$) (see **Methods**) relative to normal distribution of Archean S-MIF data (typically referred to as Archean Reference Arrays, or ARAs) coincide in several discrete stratigraphic intervals with negative excursions in carbon isotope values suggestive of methanotrophic inputs of organic carbon (Zerkle et al., 2012; Farquhar et al., 2013; Izon et al., 2015; 2017, Williford et al., 2016). The most recent of these studies (Izon et al., 2017) used available radiometric constraints to estimate the duration of a single haze event at ~1 million years. Through accelerated H₂ loss to space, the haze events are postulated by these authors to have hastened planetary oxidation. Atmospheric models of Neoproterozoic hazes suggest that CH₄ concentrations may have been at the level of 20-30 ppm (e.g. Zerkle et al., 2012), as opposed to ~2 ppm in the modern atmosphere.

Photochemical experiments with SO₂ have shown that isotope effects and resulting sulfur isotope fractionations depend on both reaction condition and the wavelength of incident radiation. Recent results (Whitehill et al. 2012; Endo et al., 2016) show that shallow positive slopes for $\Delta^{33}\text{S}/\delta^{34}\text{S}$ (<0.1) are produced by photolysis reactions at wavelengths shorter than ~220 nm, and

steeper slopes (>3) are produced when longer wavelengths (240 – 340 nm) cause photoexcitation. Similarly, photolysis of SO_2 yields steeper negative $\Delta^{36}\text{S}/\Delta^{33}\text{S}$ (<-2), while photoexcitation reactions yield positive $\Delta^{36}\text{S}/\Delta^{33}\text{S}$ values (e.g., Whitehill et al., 2012). Applying these observations to the Neoproterozoic and early Paleoproterozoic atmosphere, shifts to shallower $\Delta^{33}\text{S}/\delta^{34}\text{S}$ and steeper $\Delta^{36}\text{S}/\Delta^{33}\text{S}$ slopes would be anticipated with the transmission of shorter-wavelength UV light resulting in greater proportional photolysis reactions. This may result from the shielding of UV wavelengths by a photochemical CH_4 haze (e.g. Wolf and Toon, 2010), although quantitative details are lacking.

Pertinent to our model, it appears possible that anoxygenic photosynthesis may have been dominant during intervals when a CH_4 haze blanketed the planet. The presence of a methane haze would have shielded surface environments from harmful UV radiation that might then allow ammonia (a strong greenhouse gas) to accumulate in the lower atmosphere, and thus have shifted upward the peak wavelength of solar radiation reaching the oceans (Sagan and Chyba, 1996; Wolf and Toon, 2010). Insofar as a Neoproterozoic CH_4 haze would scatter and absorb UV and visible wavelengths (Wolf and Toon, 2010), it may have influenced photosynthetic reactions and primary productivity. In this regard, the pigments used during anoxygenic photosynthesis differ from those associated with oxygenic photosynthesis in their molecular details and peak wavelength of light absorbed. Notably the peak absorption wavelength for chlorophyll-a used by oxygen-producing plants and cyanobacteria is 100-200 nm less than the peak absorption wavelength used by bacteriochlorophylls during anoxygenic photosynthesis. The development of a significant methane haze may thus have shifted primary production of organic matter towards anoxygenic photosynthesis where the electrons necessary to drive the metabolic activity ultimately come from the oxidation of H_2S or Fe (II) (rather than from H_2O in oxygenic photosynthesis). If correct, the

dominance of the anaerobic metabolism over oxygenic photosynthesis would have decoupled inputs of R and O₂, matching a key assumption of our model.

While our resource-driven ecological model is neither calibrated for amount nor time, the consistent, robust periodic dynamics observed in the system under some parameter regimes nonetheless accord with the episodic release of CH₄ from surface environments assuming a constant, but not overwhelming, source of O₂. According to our ecological model, if this oxidant was globally distributed at low concentrations, and if resources were high, a CH₄ haze would likely result. Izon et al. (2017) suggest that the coupled carbon/sulfur isotope anomalies were associated with biospheric responses to increased nutrients, with the CH₄ flux controlled by the availability of organic carbon and sulfate. Increased loading of the oceans with nutrients (e.g., nitrogen, phosphorus) is most often associated with uplift, mountain building and weathering of exposed continental masses, all of which could erode, disrupt, or accelerate sedimentation on marginal marine platforms. In the relatively continuous sedimentary successions where these episodic coupled anomalies have been recorded, there is no physical evidence for any temporal sedimentary disruption in the form of unconformities or major hiatal surfaces. As an alternative hypothesis, we propose that if nutrients (including ferrous iron, which would have been soluble in the ocean under anoxic conditions) were widely available in the Neoarchean ocean, then the dynamic release of CH₄ and the episodic development of CH₄ hazes in the atmosphere could simply have been the result of constant low level (or gradually increasing) inputs of photosynthetic O₂ to anaerobic environments where overflow metabolism was dominant.

Geochemically documented CH₄ haze events are currently known only from the Neoarchean, but the episodic and potentially globe-encompassing ice ages of the succeeding early Paleoproterozoic Era may similarly be understood in terms of fluctuations of O₂ and greenhouse

gases in the ancient atmosphere (Pavlov et al., 2000; Bekker et al., 2005; Kopp et al., 2005; Goldblatt et al., 2006; Claire et al., 2006; Bekker and Kaufman, 2007). Insofar as the early Paleoproterozoic glacial epoch is related to the transition from the anoxic CH₄-rich atmosphere to an oxic CO₂-rich atmosphere, the multiplicity of ice ages (including at least three discrete events) in this interval is arguably related to oscillations in O₂ and CH₄ production and utilization. Related to the dynamic ecological model presented here, the rise of photosynthetic O₂ prior to each of these Paleoproterozoic glaciations would have progressively drawn down atmospheric CH₄ levels resulting in surface refrigeration. Primary productivity would then be severely limited beneath ice-covered oceans, thereby constraining the production of both fermentative products (FP) and O₂. Atmospheric CO₂ levels during the glaciations likely rose due to unabated volcanic fluxes and limited chemical weathering of exposed continental rocks (e.g., Hoffman et al., 1998). However, once threshold concentrations of CO₂ were reached, the glaciations would have ended rapidly, resulting in high rates of silicate weathering on land, and the enhanced delivery of sediments, oxidants (including sulfate and nitrate), nutrients, and alkalinity to the oceans. Melting glaciers and sea ice coupled with higher post-glacial temperatures would have flooded continental margins and stimulated both oxygenic photosynthesis and methanogenesis.

Support for such an environmental scenario comes from carbon isotope analyses of drill core samples collected from across the oldest recorded ice age deposit (the Ramsay Lake diamictite) in southern Ontario, which reveal strongly negative $\delta^{13}\text{C}$ values (with a nadir of ca. -41‰) in organic matter from sediments that accumulated in the glacial aftermath (Bekker and Kaufman, 2007). The magnitude and duration of ^{13}C depletion in the post-glacial Pecors Formation (expressed over 100 meters of sedimentary strata) is consistent with long-term methanogenic production and methanotrophic oxidation of CH₄ in a shallow oceanic setting. Multiple rare sulfur

isotope abundances in pyrite extracted from the same Pecors samples reveal even more profound negative deviations in $\Delta^{36}\text{S}/\Delta^{33}\text{S}$ values relative to the ARA (Wing et al., 2002; 2004), which may similarly record atmospheric processes during this (and other) post-glacial early Paleoproterozoic CH_4 haze events. Considering our model results, we suggest that the end of the early Paleoproterozoic glacial epoch around 2.2 billion years ago likely marked the end of extended methane-dominated atmospheres of the early Earth. This repetitive series of ice ages was followed by a remarkably long-lived (>100 million years) interval of highly positive carbon isotope compositions in carbonates known as the Lomagundi-Jatuli event (Karhu and Holland, 1996; Bachan and Kump, 2015; Mänd et al., 2020). Insofar as significant ^{13}C enrichments in marine carbonates are driven by the burial and sequestration of ^{12}C -rich organic matter (Broecker, 1970; Hayes, 1983), the attendant photosynthetic production and release of O_2 at this time would most likely have marked the transition to an oxygenated atmosphere with carbon dioxide acting as the major greenhouse gas thereafter.

Episodically enhanced CH_4 release and buildup in the atmosphere later in Earth history may be related to the tectonic reorganization of the mobile continents, which modulate the delivery of reduced gases from spreading ridges and weathering fluxes from uplifted mountains. These variable fluxes would have impacted the photosynthetic production of organic matter and degree of ocean oxygenation during discrete climatic and biological events. Insofar as our model results indicate that these factors control CH_4 fluxes to the atmosphere, we speculate that younger ice ages of the Neoproterozoic and Paleozoic eras, as well as some Phanerozoic mass extinctions, may be linked to the dynamic microbial processes modeled here.

Conclusion

Patterns of CH₄ production and release in both driven and non-driven versions of a novel microbial ecological model, based upon the metabolic repercussions of the Warburg Effect, suggest that resources and O₂ play important roles in defining the dynamics of microbial ecosystems over both short and long-time scales. We demonstrate that forced low level inputs of O₂ result in pulsed releases of CH₄ via overflow metabolism, a pattern that may explain periodic CH₄ haze events and episodic glaciation deep in Earth's history. Such pulsed methane releases also occur in modern-day wetlands where tidal flooding repeatedly tips the balance between oxic and anoxic conditions (see Zhu et al., 2017), suggesting similar mechanisms may be at work. The qualitative behavior of this ecosystem model provides a platform for understanding the dynamics of mixed microbial communities in ancient, modern, and future ecosystems. More detailed models that blend empirical gas-release data with calibrated input parameters should shed further light on the links between microbial communities and the timing of atmospheric dynamics.

Methods

Modeling Microbial System Dynamics

We model the dynamics of the interacting microbial system as a system of ordinary differential equations (ODEs). First, we define functions pertaining to the Warburg Effect (eq. 1-2) and then describe the equations for each of the microbial functional groups.

We begin by defining two functions that describe Warburg Effect dynamics in aerobic heterotrophic bacteria (A):

$$f([R], A) = \begin{cases} \alpha A ([R] - R_0) & \text{for } [R] > R_0 \\ 0 & \text{for } [R] < R_0 \end{cases} \quad \text{Eq. 1 (overflow metabolism of A)}$$

$$g([R], A) = \begin{cases} \varepsilon A R_1 & \text{for } [R] > R_1 \\ \varepsilon A [R] & \text{for } [R] < R_1 \end{cases} \quad \text{Eq. 2 (respiratory function of A)}$$

The first function $f(R, A)$ denotes the process of overflow metabolism, where, despite the presence of ample oxygen, heterotrophic aerobic bacteria (A) produce fermentation products when resources are abundant. In this function, at resource concentrations higher than base levels (represented by R_0), there is a linear increase in overflow products (see Basan et al., 2015, Swain and Fagan, 2018). The second function $g(R, A)$ denotes the respiratory function (see Eq. 3 below) for these organisms (with the maximal resource usage condition represented by resource concentration of R_1). Both α and ε are proportionality constants that denote the efficiency of overflow metabolism and respiration, respectively, and fall in the range $[0, 1]$.

The differential equations below pertain to changes through time of the dominance of metabolic functional groups as the different microbial functional groups (A, B, C, F, and G as defined in **Fig. 3.1**) compete for available substrates under variable environmental conditions.

$$\frac{dA}{dt} = (1 - \mu) [O_2] g([R], A) - s A \quad \text{Eq. 3}$$

$$\frac{dB}{dt} = (1 - \kappa) \beta [FP] B - u B \quad \text{Eq. 4}$$

$$\frac{dC}{dt} = (1 - \lambda) \gamma [O_2][CH_4] C - v C \quad \text{Eq. 5}$$

$$\frac{dF}{dt} = (1 - \xi) \omega [R] F - a_1 F \quad \text{Eq. 6}$$

$$\frac{dG}{dt} = (1 - \psi) \Omega [CH_4] G - a_2 G \quad \text{Eq. 7}$$

$$\frac{dH}{dt} = (1 - \rho) \tau [FP] H - m H \quad \text{Eq. 8}$$

The first term on the right-hand side (RHS) of each differential equation is positive and denotes the process of assimilation of substrates needed for growth (increase in biomass) and energy production for the metabolic class. The second RHS term in each equation is negative and denotes

the energy requirement for homeostasis (see **Table 3.1**). In equations 3-8, the terms governing resource incorporation efficiency, generically denoted $(1 - X)$ with X replaced by the consortium-specific parameters $(\mu, \kappa, \lambda, \xi \text{ and } \psi)$, specify the proportion of the required substrate in the extracellular environment that is incorporated into cells, which can then be used for making biomass by those microbes. Variation in these efficiencies is assumed to be modulated by biochemical and physical constraints of the different microbial functional groups. The remaining reservoir of resources outside of the cell either leave the system as gases or are available for later incorporation. All twelve coefficients of efficiency $(\mu, \kappa, \lambda, \xi, \psi, \rho, \varepsilon, \beta, \gamma, \omega, \Omega, \text{ and } \tau)$ lie in the range $[0, 1]$. The coefficients related to homeostasis $(s, u, v, a_1, a_2, \text{ or } m)$ represent the basal metabolic requirement values for microbial functional groups A, B, C, F, G, and H respectively. All parameters and their interpretations are summarized in **Table 3.1**.

Functional group	Resource incorporation efficiency	Resource utilization efficiency	Energy requirements for homeostasis
A (aerobic heterotrophs)	$(1 - \mu)$	ε (for R)	s
B (methanogens)	$(1 - \kappa)$	β (for FP)	u
C (aerobic methanotrophs)	$(1 - \lambda)$	γ (for O ₂ and CH ₄)	v
F (fermenters)	$(1 - \xi)$	ω (for R)	a_1
G (anaerobic methanotrophs)	$(1 - \psi)$	Ω (for CH ₄)	a_2
H (anaerobic heterotrophs)	$(1 - \rho)$	τ (for FP)	m

Table 3.1: Differential equation parameters governing the dynamics of a system of five microbial functional groups. Parameters summarize the efficiency of resource incorporation, the efficiency of resource utilization, and the energy requirements for homeostasis.

The following set of differential equations reflect dynamic changes in the abundance of resources or metabolic products in the five microbial functional groups through time:

$$\frac{d[R]}{dt} = -f([R], A) - g([R], A) - \xi \omega [R] F + \sigma_1(t) \quad \text{Eq 9}$$

$$\frac{d[FP]}{dt} = f([R], A) - \beta B [FP] + \xi \omega R F - \tau[FP]H \quad \text{Eq 10}$$

$$\frac{d[CO_2]}{dt} = \mu [O_2] g([R], A) + \gamma \lambda [O_2] [CH_4] C + \psi \Omega [CH_4] G + \rho \tau[FP]H \quad \text{Eq 11}$$

$$\frac{d[CH_4]}{dt} = \kappa \beta B [FP] - \gamma [O_2] [CH_4] C - \Omega [CH_4] G \quad \text{Eq 12}$$

$$\frac{d[O_2]}{dt} = -\{\mu [O_2] g([R], A) + \gamma \lambda [O_2] [CH_4] C\} + \sigma_2(t) \quad \text{Eq 13}$$

In equations 8-12 the RHS positive terms indicate either input or production of the resources, while negative terms indicate metabolic resource consumption. All the terms involve interaction of the resources with one or more of the microbial functional groups, which are modelled in equations 3-8, except for the ‘forcing terms’ $\sigma_1(t)$ and $\sigma_2(t)$. These terms (see Eqs. 9 and 13) denote the time-dependent input rates of R and O_2 that are controlled externally. In this model, CO_2 is considered a sink that evades to the atmosphere and is decoupled from the autotrophic processes. This assumption makes R an independent variable, and further ensures that substrate production is not photoautotrophic (the case in many scenarios). Moreover, as developed here, the ecological model is not spatially compartmented in order to make the results more general in nature.

There exist many methanogenic systems where oxic and anoxic parts are heterogeneously mixed, albeit at generally small scales or temporally, such as in wetlands (Kirsche et al., 2013). In this ecological model we have simplified the process of methanogenesis in several ways: i) FPs can be one of many substrates, including formate, acetate, and lactate, all of which can be used for methane production by various methanogens; and ii) for hydrogen (H_2) + CO_2 methane production,

we assume that reduced reactants are unlimited, and insofar as CO_2 is a sink and evades the system as a gas, it is unlimited as well (i.e., it is not used for autotrophic processes that deliver new resources to the community). For the anaerobic oxidation of methane by functional group G, we lump all of the potential anaerobic methanotrophs here, but recognizing that both CO_2 and carbonate alkalinity may be formed, we include a conversion efficiency term (η) to calculate the relative concentrations of CO_2 alone.

This coupled differential equation system was solved through Wolfram Mathematica v. 12.1 and R v. 3.6.2. A regular parameter sweep was first performed for the non-driven case (no external input: $\sigma_1, \sigma_1 = 0$) using Latin Hypercube sampling (LHS) for 500,000 different scenarios using R packages *deSolve* (Soetaert et al., 2010) and *lhs* (Carnell, 2020). We then calculated the partial rank correlation coefficients (PRCCs) to explore the effect of each parameter and initial condition on the concentrations of gases (CH_4 , O_2 and CO_2) through R package *sensitivity* (Iooss et al., 2020).

To further explore the trends in these gas emissions, we calculated the relative proportions of these gases under a set of base parameters that depict biologically possible scenarios (according to data from Payne and Wiebe, 1978; Geyer et al., 2016; Bowles et al., 2019; and Morris and Schmidt, 2013). All the efficiencies were set at half maximum for the base case. We then varied one parameter at a time to look at the dependence on that specific variable on gas emissions as a function of resource and oxygen initial conditions. In doing so, we collated the data for different initial conditions (such as different initial ratios of the microbial functional groups) and plotted the averaged values of the equilibrium state as a function of initial resource and oxygen inputs. For driven dynamics, we followed the same procedure of parameter dependence, but now including the two R and O_2 external input parameters, $\sigma_1(t)$ and $\sigma_2(t)$. To calculate the period and amplitude

of the driven dynamics, values were numerically estimated based on dynamic results from the first few oscillations. Periodicity statistics for the dynamics resulting from both high and low extreme values of σ_2 are estimates given the difficulty of estimating them for very long or extremely short oscillations that occur at these values of $\sigma_2(t)$.

For looking at the sensitivity of the model to addition or removal of microbial functional groups, we ran the same tests for a model without H (anaerobic heterotrophs). To measure sensitivity of the demarcation of the oscillatory phase in the space of resource and oxygen addition (σ_1 - σ_2 space), across a wide variety of parameter combinations, we first used LHS to sample 100,000 combinations through the parameters space. We calculated, for each set of parameters, twenty points on the σ_1 - σ_2 space, estimated the relative value for which there is a phase change from oscillatory behavior to no oscillations, and from these values calculated the slope of the line which demarcates the space for each set of LHS parameters. We performed PRCC analysis to find the dependence of the slope value on the parameters.

Notation for Sulfur Isotope Compositions

Sulfur-isotope ratios are conventionally reported in delta (δ) notation and reflect the permil (‰) deviation of the ratio of the least abundant isotope (^{33}S , ^{34}S , or ^{36}S) to the most abundant isotope (^{32}S), relative to the same ratio in an international reference standard (Vienna Canyon Diablo Troilite, V-CDT). For example, $\delta^{34}\text{S} = [({}^{34}\text{S}/{}^{32}\text{S})_{\text{sample}}/({}^{34}\text{S}/{}^{32}\text{S})_{\text{V-CDT}} - 1] \times 1000$. The majority of terrestrial processes fractionate S-isotopes mass-dependently, whereby $\delta^{33}\text{S} \approx 0.515 * \delta^{34}\text{S}$ and $\delta^{36}\text{S} \approx 1.91 * \delta^{34}\text{S}$. Departure from mass-dependent behavior, or mass-independent fractionation (S-MIF), is expressed in capital-delta (Δ) notation as either non-zero $\Delta^{33}\text{S} =$

$$[(^{33}\text{S}/^{32}\text{S})_{\text{sample}}/(^{33}\text{S}/^{32}\text{S})_{\text{V-CDT}} - (^{34}\text{S}/^{32}\text{S})_{\text{sample}}/(^{34}\text{S}/^{32}\text{S})_{\text{V-CDT}}]^{0.515} \text{ or } \Delta^{36}\text{S} =$$

$$[(^{36}\text{S}/^{32}\text{S})_{\text{sample}}/(^{36}\text{S}/^{32}\text{S})_{\text{V-CDT}} - (^{34}\text{S}/^{32}\text{S})_{\text{sample}}/(^{34}\text{S}/^{32}\text{S})_{\text{V-CDT}}]^{1.9}.$$

Acknowledgements

This research was supported in part by NSF award EAR-1338810 to AJK. AS would like to thank the COMBINE program at the University of Maryland (DGE-1632976) for technical and academic support.

Code and data availability

All the codes that were used to explore the ODE model and generate the relevant figures are available publicly at the following github repository:

https://github.com/anshuman21111/ch4_haze .

All extra figures and simulation data, above and exceeding the supplementary materials are also found on this repository.

References

1. Anbar, A.D. and Knoll, A.H., 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge?. *science*, 297(5584), pp.1137-1142.
2. Bachan, A. and Kump, L.R., 2015. The rise of oxygen and siderite oxidation during the Lomagundi Event. *Proceedings of the National Academy of Sciences*, 112(21), pp.6562-6567.
3. Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J.R. and Hwa, T., 2015. Overflow metabolism in *Escherichia coli* results from efficient proteome allocation. *Nature*, 528(7580), pp.99-104.
4. Bekker, A., Kaufman, A.J., Karhu, J.A. and Eriksson, K.A., 2005. Evidence for Paleoproterozoic cap carbonates in North America. *Precambrian Research*, 137(3-4), pp.167-206.
5. Bekker, A. and Kaufman, A.J., 2007. Oxidative forcing of global climate change: a biogeochemical record across the oldest Paleoproterozoic ice age in North America. *Earth and Planetary Science Letters*, 258(3-4), pp.486-499.
6. Carlson, R.P., Beck, A.E., Phalak, P., Fields, M.W., Gedeon, T., Hanley, L., Harcombe, W.R., Henson, M.A. and Heys, J.J., 2018. Competitive resource allocation to metabolic pathways contributes to overflow metabolisms and emergent properties in cross-feeding microbial consortia. *Biochemical Society Transactions*, 46(2), pp.269-284.
7. Carnell, R., 2020. lhs: Latin Hypercube Samples. R package version 1.0.2.
<https://CRAN.R-project.org/package=lhs>
8. Chistoserdova, L., Vorholt, J.A. and Lidstrom, M.E., 2005. A genomic view of methane oxidation by aerobic bacteria and anaerobic archaea. *Genome biology*, 6(2), p.208.

9. Eigenbrode, J.L. and Freeman, K.H., 2006. Late Archean rise of aerobic microbial ecosystems. *Proceedings of the National Academy of Sciences*, 103(43), pp.15759-15764.
10. Farquhar, J., Cliff, J., Zerkle, A.L., Kamysny, A., Poulton, S.W., Claire, M., Adams, D. and Harms, B., 2013. Pathways for Neoarchean pyrite formation constrained by mass-independent sulfur isotopes. *Proceedings of the National Academy of Sciences*, 110(44), pp.17638-17643.
11. Halverson, G.P., Hoffman, P.F., Schrag, D.P. and Kaufman, A.J., 2002. A major perturbation of the carbon cycle before the Ghaub glaciation (Neoproterozoic) in Namibia: Prelude to snowball Earth?. *Geochemistry, Geophysics, Geosystems*, Vol. 3(6)
12. Hayes, J.M., 1994. Early life on Earth. In *Nobel symposium* (Vol. 84, pp. 220-36).
13. Iooss, B., Da Veiga, S., Janon, A., Pujol, G., et al., 2020. sensitivity: Global Sensitivity Analysis of Model Outputs. R package version 1.20.0. <https://CRAN.R-project.org/package=sensitivity>
14. Izon, G., Zerkle, A.L., Zhelezinskaia, I., Farquhar, J., Newton, R.J., Poulton, S.W., Eigenbrode, J.L. and Claire, M.W., 2015. Multiple oscillations in Neoarchean atmospheric chemistry. *Earth and Planetary Science Letters*, 431, pp.264-273.
15. Izon, G., Zerkle, A.L., Williford, K.H., Farquhar, J., Poulton, S.W. and Claire, M.W., 2017. Biological regulation of atmospheric chemistry en route to planetary oxygenation. *Proceedings of the National Academy of Sciences*, 114(13), pp.E2571-E2579.
16. Kasting, J.F., 2001. The rise of atmospheric oxygen. *Science*, 293(5531), pp.819-820.

17. Kirschke, S., Bousquet, P., Ciais, P., Saunois, M., Canadell, J.G., Dlugokencky, E.J., Bergamaschi, P., Bergmann, D., Blake, D.R., Bruhwiler, L. and Cameron-Smith, P., 2013. Three decades of global methane sources and sinks. *Nature geoscience*, 6(10), pp.813-823.
18. Konhauser, K.O., Pecoits, E., Lalonde, S.V., Papineau, D., Nisbet, E.G., Barley, M.E., Arndt, N.T., Zahnle, K. and Kamber, B.S., 2009. Oceanic nickel depletion and a methanogen famine before the Great Oxidation Event. *Nature*, 458(7239), pp.750-753.
19. Kopp, R.E., Kirschvink, J.L., Hilburn, I.A. and Nash, C.Z., 2005. The Paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. *Proceedings of the National Academy of Sciences*, 102(32), pp.11131-11136.
20. Kreft, J.U., Griffin, B.M. and González-Cabaleiro, R., 2020. Evolutionary causes and consequences of metabolic division of labour: why anaerobes do and aerobes don't. *Current opinion in biotechnology*, 62, pp.80-87.
21. Lyu, Z., Shao, N., Akinyemi, T. and Whitman, W.B., 2018. Methanogenesis. *Current Biology*, 28(13), pp.R727-R732.
22. Manzoni, S., Taylor, P., Richter, A., Porporato, A. and Ågren, G.I., 2012. Environmental and stoichiometric controls on microbial carbon use efficiency in soils. *New Phytologist*, 196(1), pp.79-91.
23. Pavlov, A.A., Kasting, J.F., Eigenbrode, J.L. and Freeman, K.H., 2001. Organic haze in Earth's early atmosphere: Source of low-¹³C Late Archean kerogens?. *Geology*, 29(11), pp.1003-1006
24. Saunois, M., Bousquet, P., Poulter, B., Peregon, A., Ciais, P., Canadell, J.G., Dlugokencky, E.J., Etiope, G., Bastviken, D., Houweling, S. and Janssens-Maenhout, G.,

2016. The global methane budget 2000–2012. *Earth System Science Data*, 8(2), pp.697-751.
25. Soetaert, K., Petzoldt, T. and Setzer, R.W., 2010. Package deSolve: solving initial value differential equations in R. *J Stat Softw*, 33(9), pp.1-25.
26. Swain, A. and Fagan, W.F., 2018. A mathematical model of the Warburg Effect: Effects of cell size, shape and substrate availability on growth and metabolism in bacteria. *Math. Biosci. Eng*, 16, pp.168-186.
27. Voet, D. and Voet, J.G., 2004. *Biochemistry*. Hoboken. John Wiley & Sons
28. Warburg, O., 1956. On the origin of cancer cells. *Science*, 123(3191), pp.309-314.
29. Ward, L.M., Rasmussen, B. and Fischer, W.W., 2019. Primary productivity was limited by electron donors prior to the advent of oxygenic photosynthesis. *Journal of Geophysical Research: Biogeosciences*, 124(2), pp.211-226.
30. Williford, K.H., Ushikubo, T., Lepot, K., Kitajima, K., Hallmann, C., Spicuzza, M.J., Kozdon, R., Eigenbrode, J.L., Summons, R.E. and Valley, J.W., 2016. Carbon and sulfur isotopic signatures of ancient life and environment at the microbial scale: Neoproterozoic shales and carbonates. *Geobiology*, 14(2), pp.105-128.
31. Wing, B.A., Bekker, A., Brabson, L., Farquhar, J., Kaufman, A.J. and Rumble, D., 2004, November. Atmospheric chemistry of SO₂ during the earliest Huronian glaciation. In *Abstr. Progr. Geol. Soc. Am (Vol. 36)*.
32. Wing, B.A., Brabson, E., Farquhar, J., Kaufman, A.J., Rumble, D. and Bekker, A., 2002, August. Delta S-33, delta S-34 and delta C-13 constraints on the Paleoproterozoic atmosphere during the earliest Huronian glaciation. In *Geochimica et Cosmochimica Acta Vol. 66, No. 15 A*, pp. A840-A840

33. Wolf, E.T. and Toon, O.B., 2010. Fractal organic hazes provided an ultraviolet shield for early Earth. *Science*, 328(5983), pp.1266-1268.
34. Yan, Z., Joshi, P., Gorski, C.A. and Ferry, J.G., 2018. A biochemical framework for anaerobic oxidation of methane driven by Fe (III)-dependent respiration. *Nature communications*, 9(1), pp.1-9.
35. Zerkle, A.L., Claire, M.W., Domagal-Goldman, S.D., Farquhar, J. and Poulton, S.W., 2012. A bistable organic-rich atmosphere on the Neoarchaeon Earth. *Nature Geoscience*, 5(5), pp.359-363.
36. Zhelezinskaia, I., Kaufman, A.J., Farquhar, J. and Cliff, J., 2014. Large sulfur isotope fractionations associated with Neoarchean microbial sulfate reduction. *Science*, 346(6210), pp.742-744.

Supplementary materials

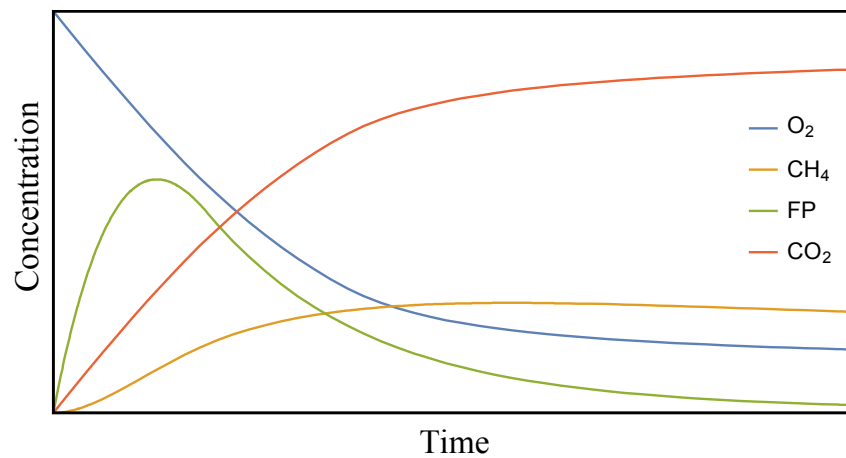


Figure S3.1: Dynamic variations in O_2 , FP , CH_4 , and CO_2 through time with set initial boundary conditions so that (relative) concentrations eventually level out when the system reaches steady-state equilibrium, for a non-driven case.

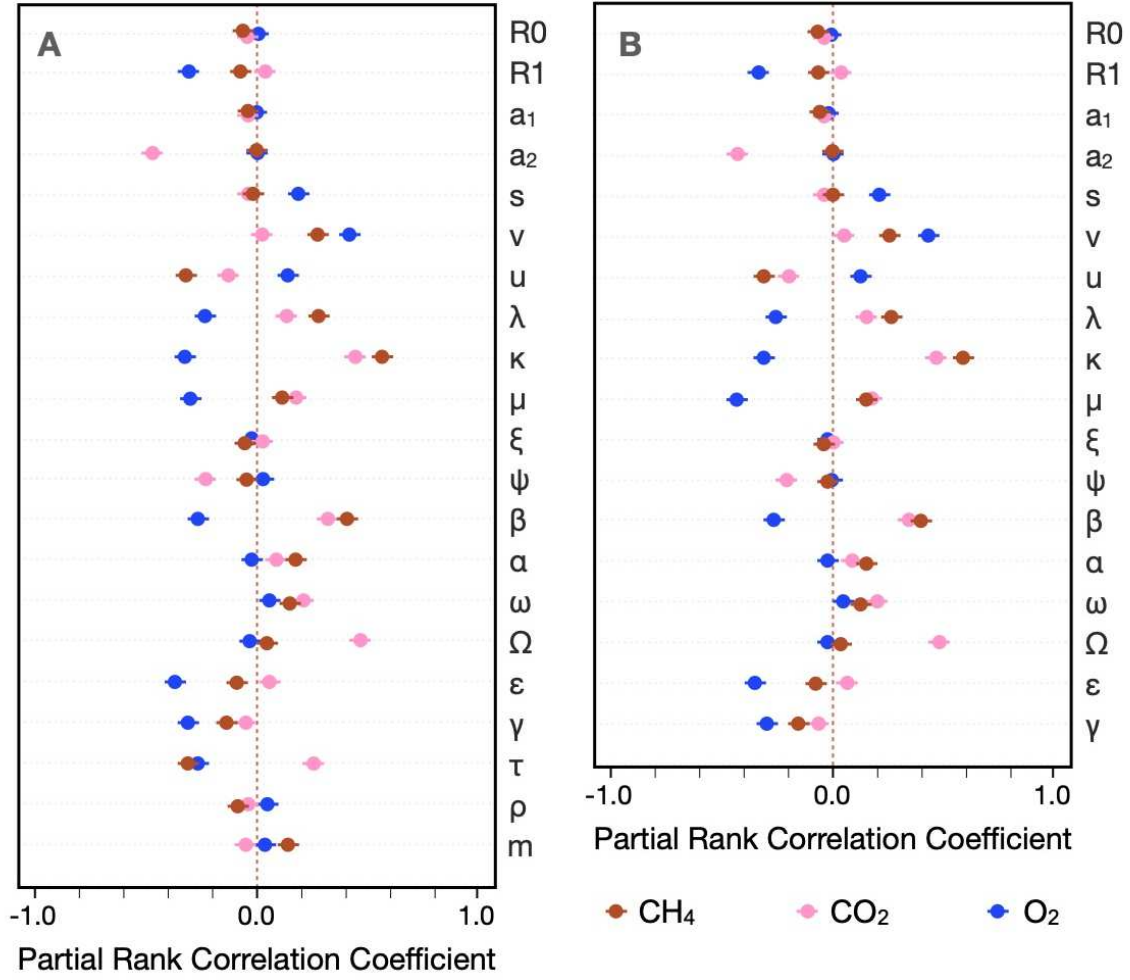


Figure S3.2: Latin hypercube sampling (LHS) based Partial rank correlation coefficient (PRCC) (which varies from -1 to 1) of the model parameters with respect to the final gas concentrations in (A) the complete model, (B) in the model with H removed (for checking sensitivity). (A) and (B) have negligible differences, signifying model capacity to perturbations and additions.

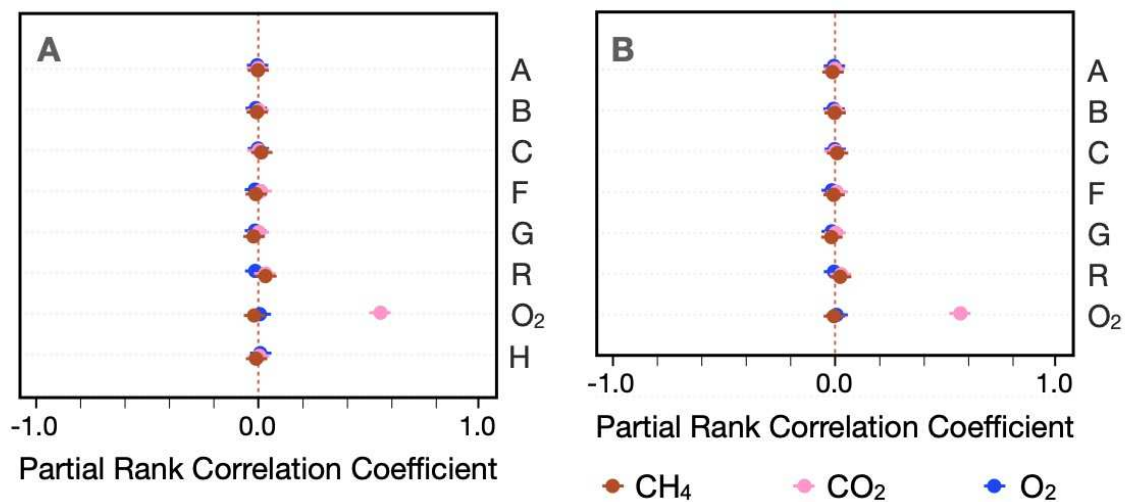


Figure S3.3: Latin hypercube sampling (LHS) based Partial rank correlation coefficient (PRCC) (which varies from -1 to 1) of the model initial conditions with respect to the final gas concentrations in (A) the complete model, (B) in the model with H removed (for checking sensitivity). (A) and (B) have negligible differences, signifying model capacity to perturbations and additions. Please note the extremely small values (absolute value <0.05) of PRCC in all cases except that of CO₂ dependence on O₂ initial value.

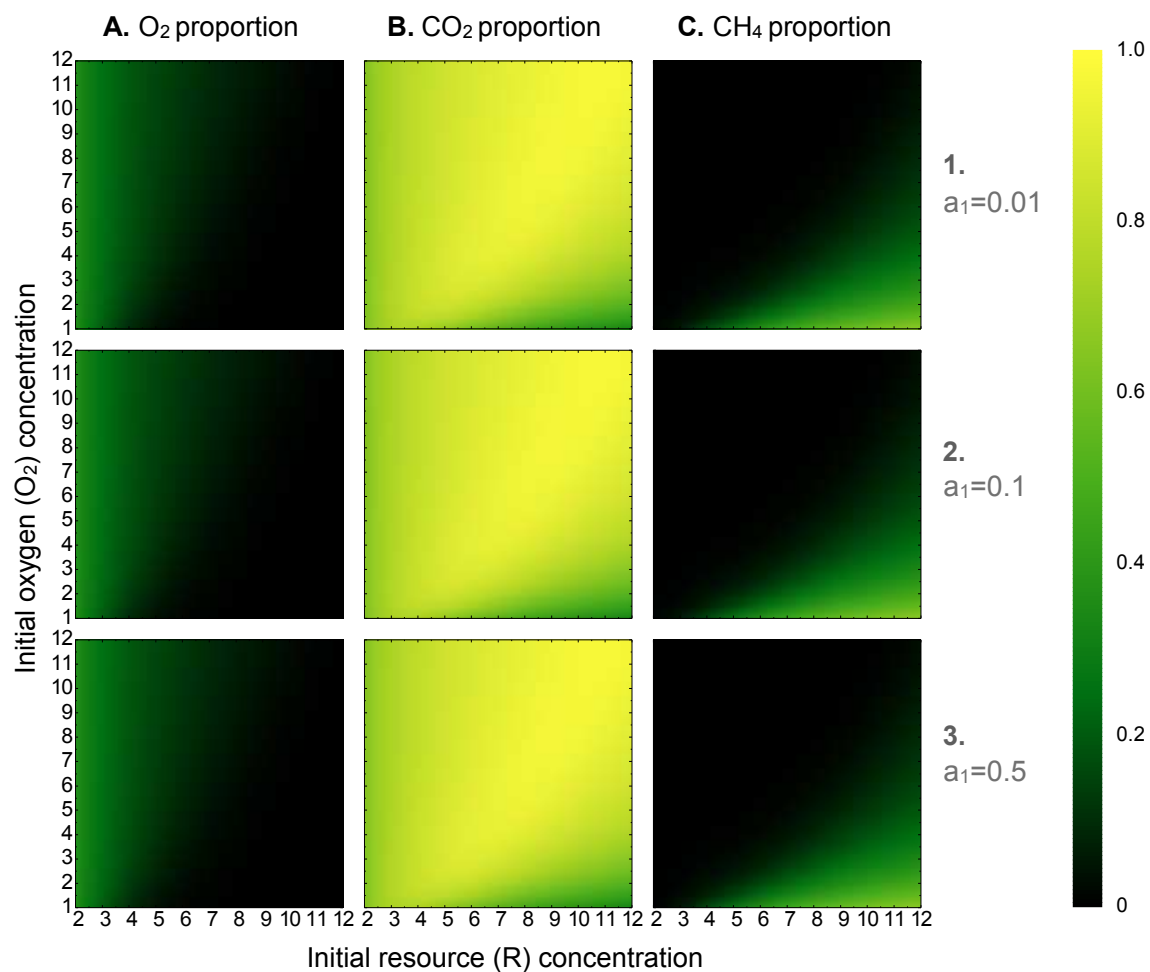


Figure S3.4A: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with the energy requirement for fermenters (a_1) set at 1, 10, and 50% in rows 1-3, respectively.

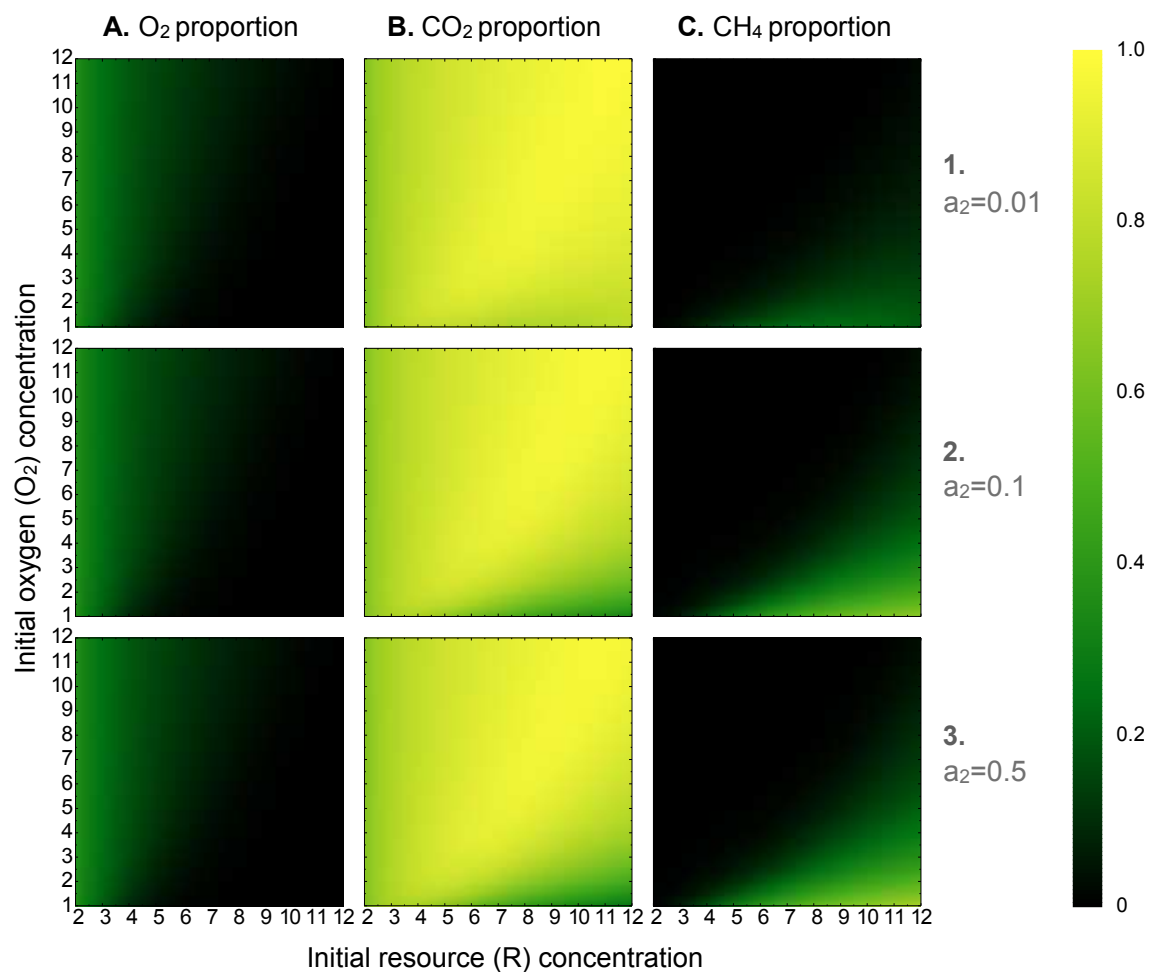


Figure S3.4B: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with the energy requirement for anaerobic methanotrophs (a_2) set at 1, 10, and 50% in rows 1-3, respectively.

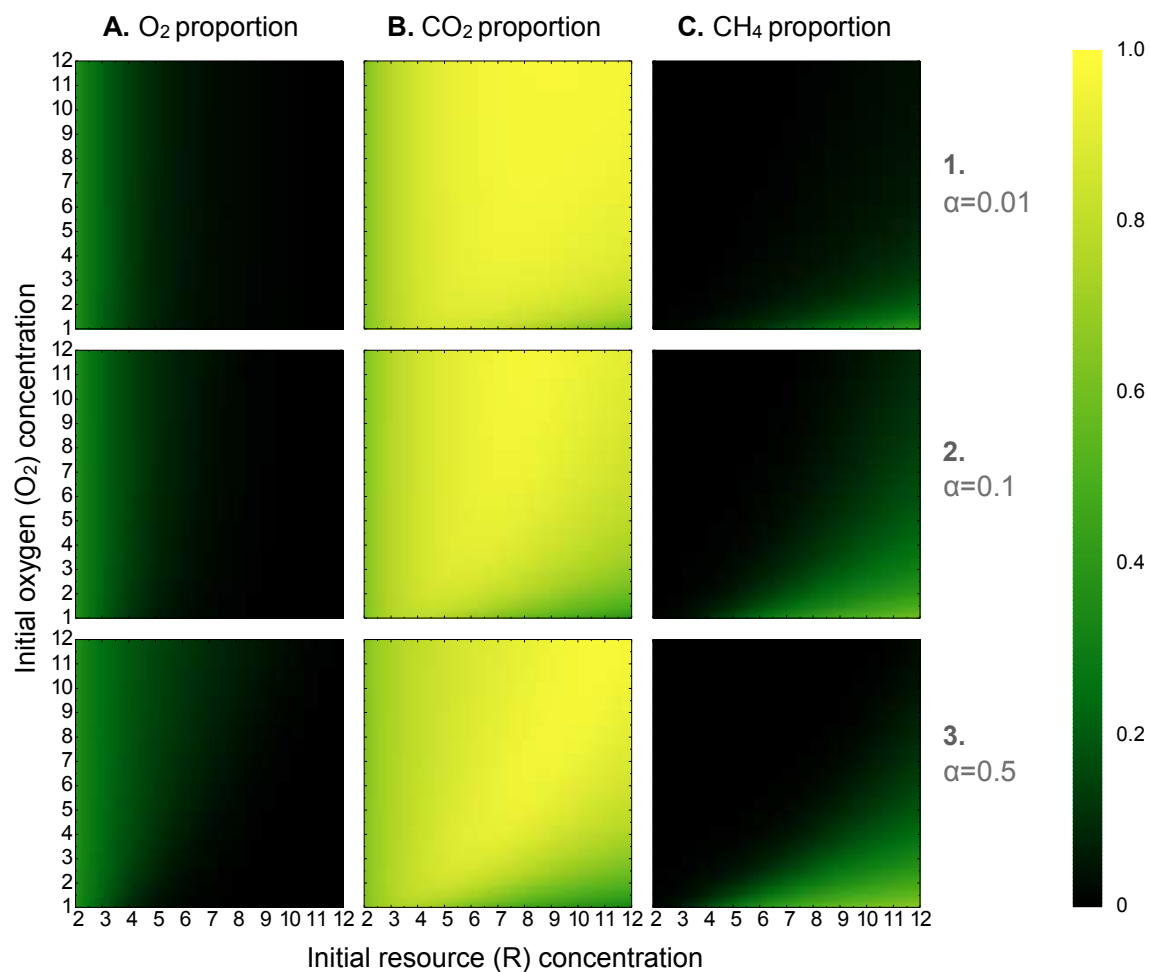


Figure S3.4C: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with efficiency of overflow metabolism in aerobic heterotrophs (α) set at 1, 10, and 50% in rows 1-3, respectively.

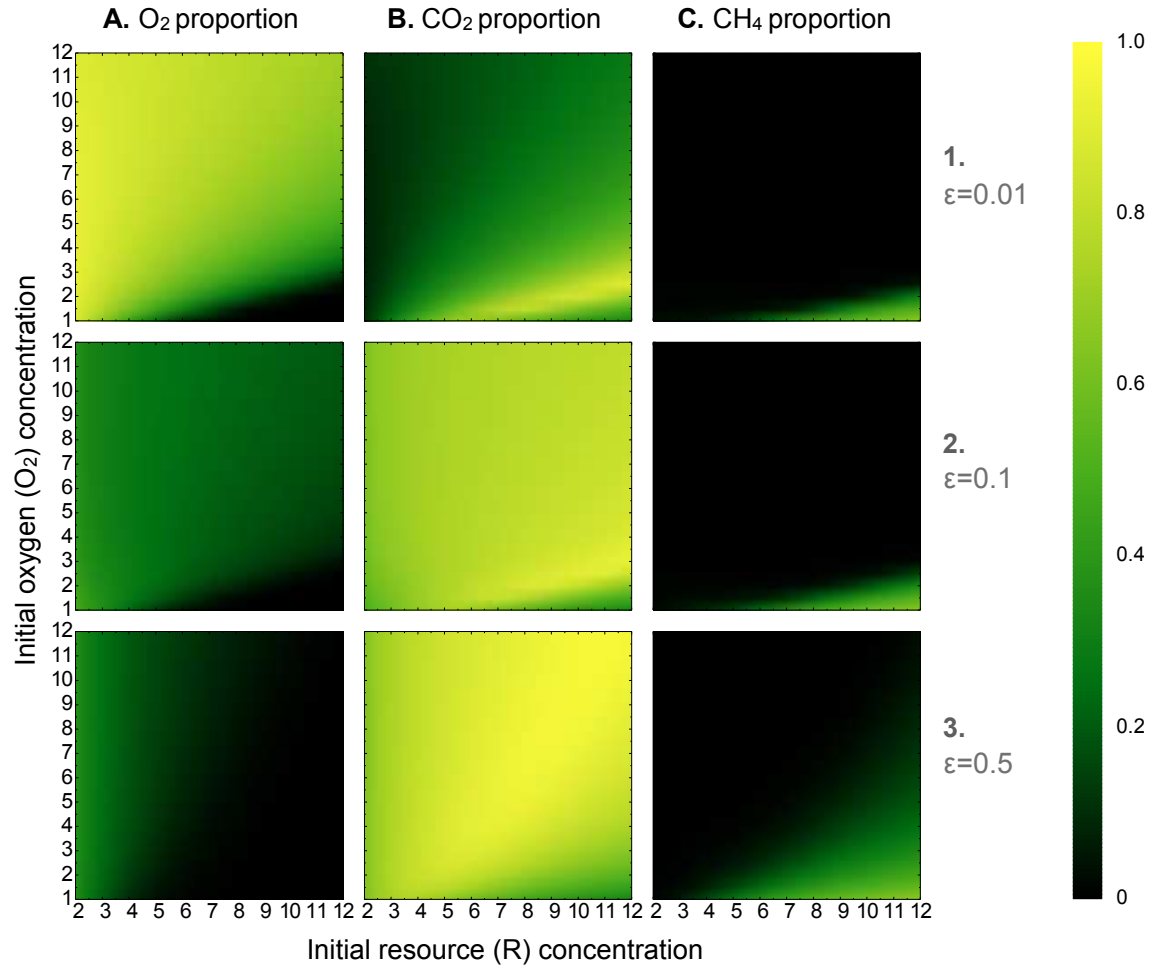


Figure S3.4D: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource utilization efficiency for respiration in aerobic heterotrophs (ϵ) set at 1, 10, and 50% in rows 1-3, respectively.

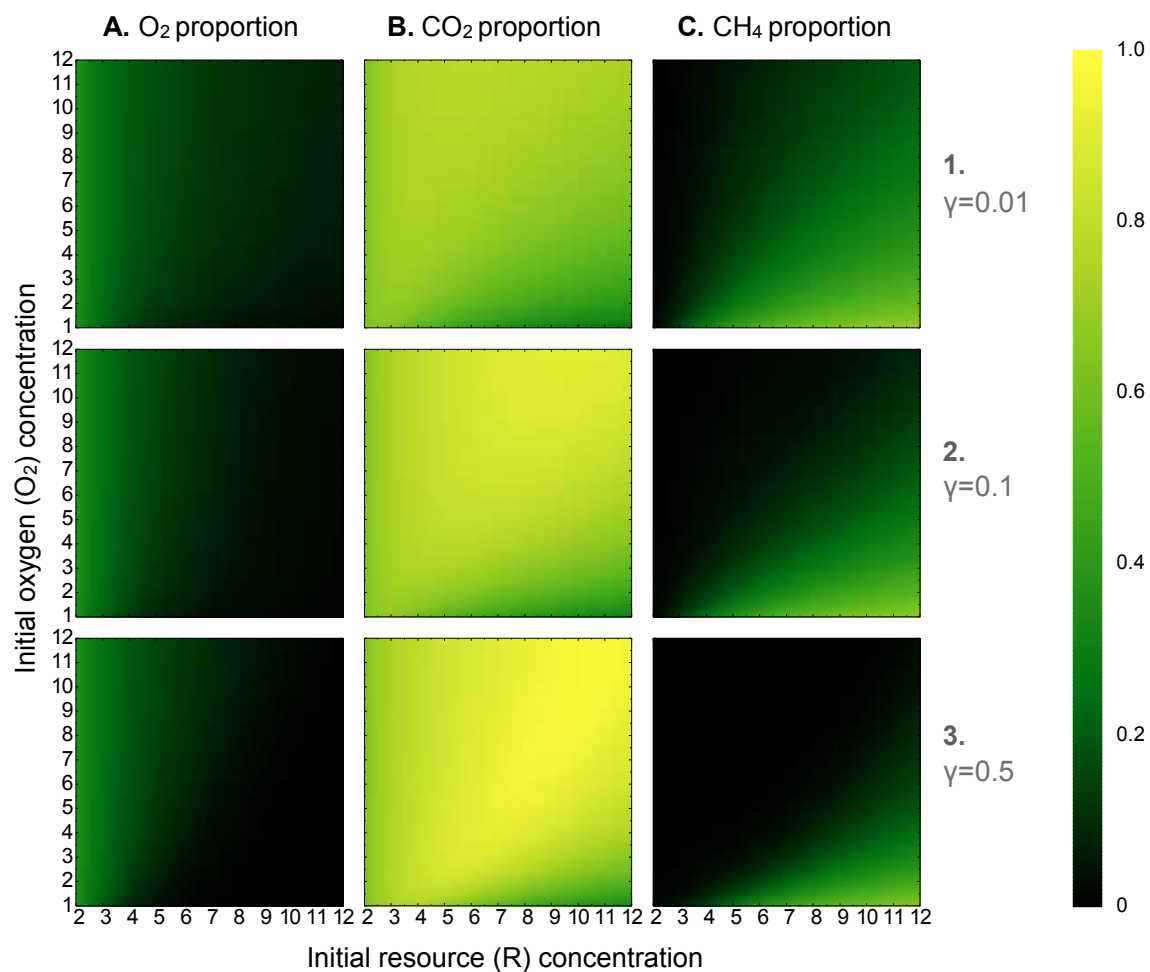


Figure S3.4E: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource utilization efficiency in aerobic methanotrophs (γ) set at 1, 10, and 50% in rows 1-3, respectively.

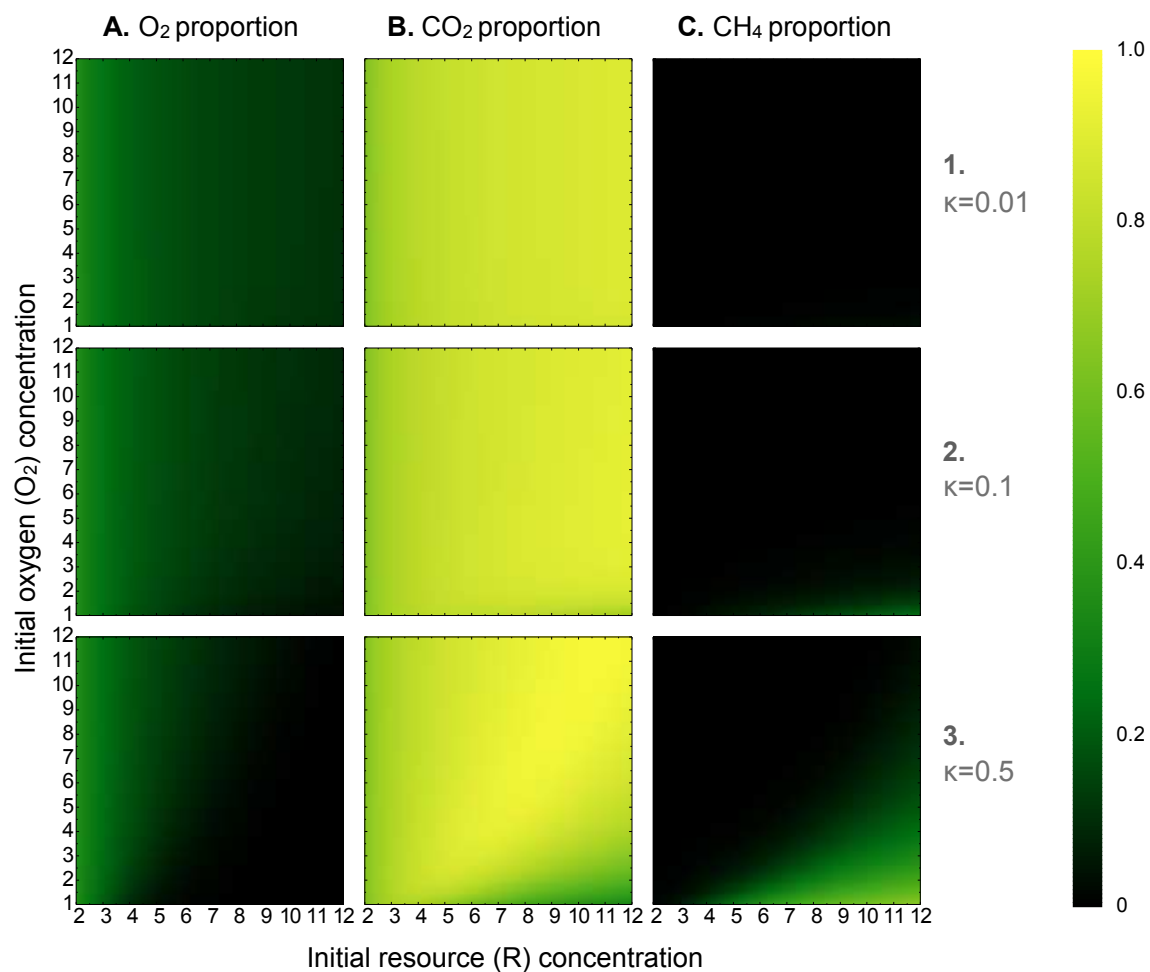


Figure S3.4F: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource incorporation efficiency in methanogens (κ) set at 1, 10, and 50% in rows 1-3, respectively.

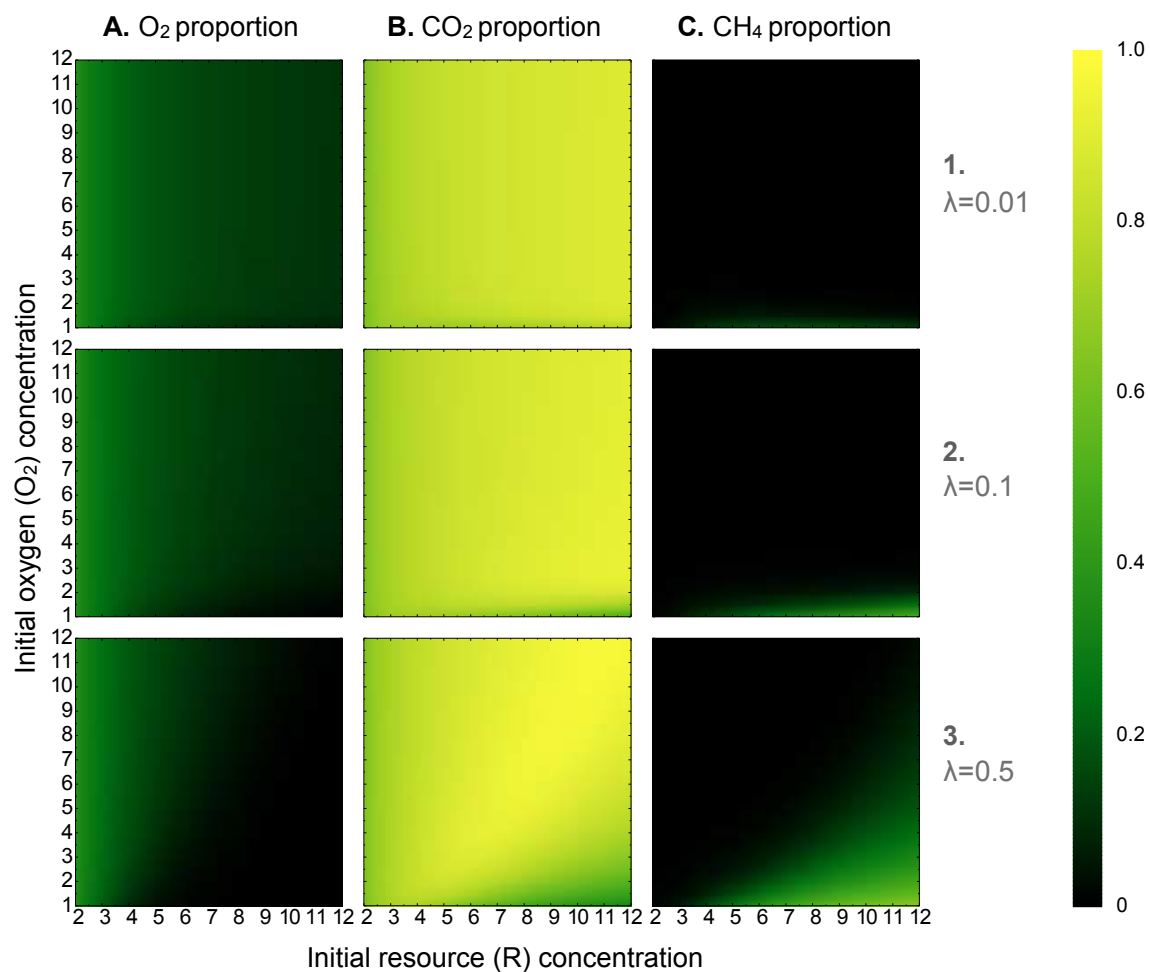


Figure S3.4G: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource incorporation efficiency in aerobic heterotrophs (λ) set at 1, 10, and 50% in rows 1-3, respectively.

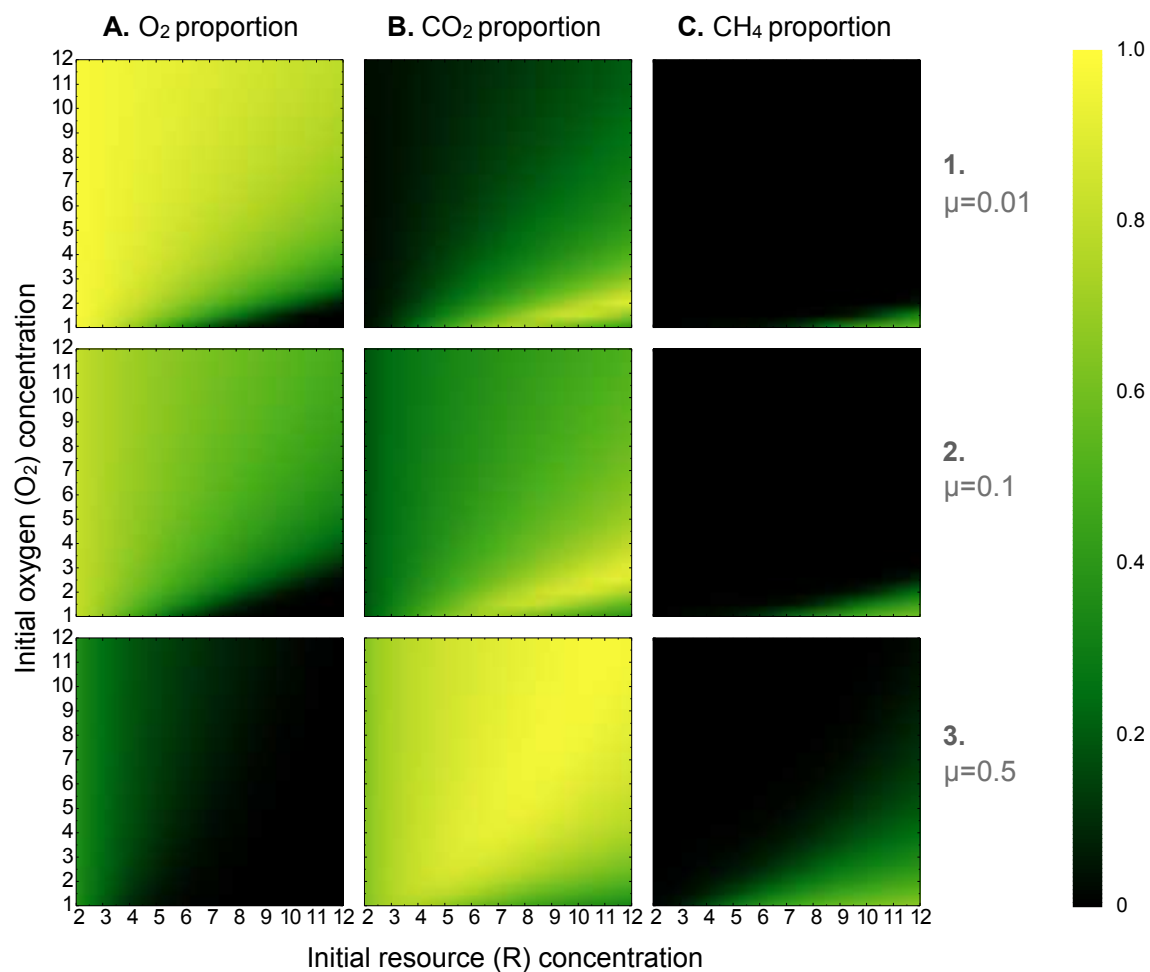


Figure S3.4H: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource incorporation efficiency in aerobic methanotrophs (μ) set at 1, 10, and 50% in rows 1-3, respectively.

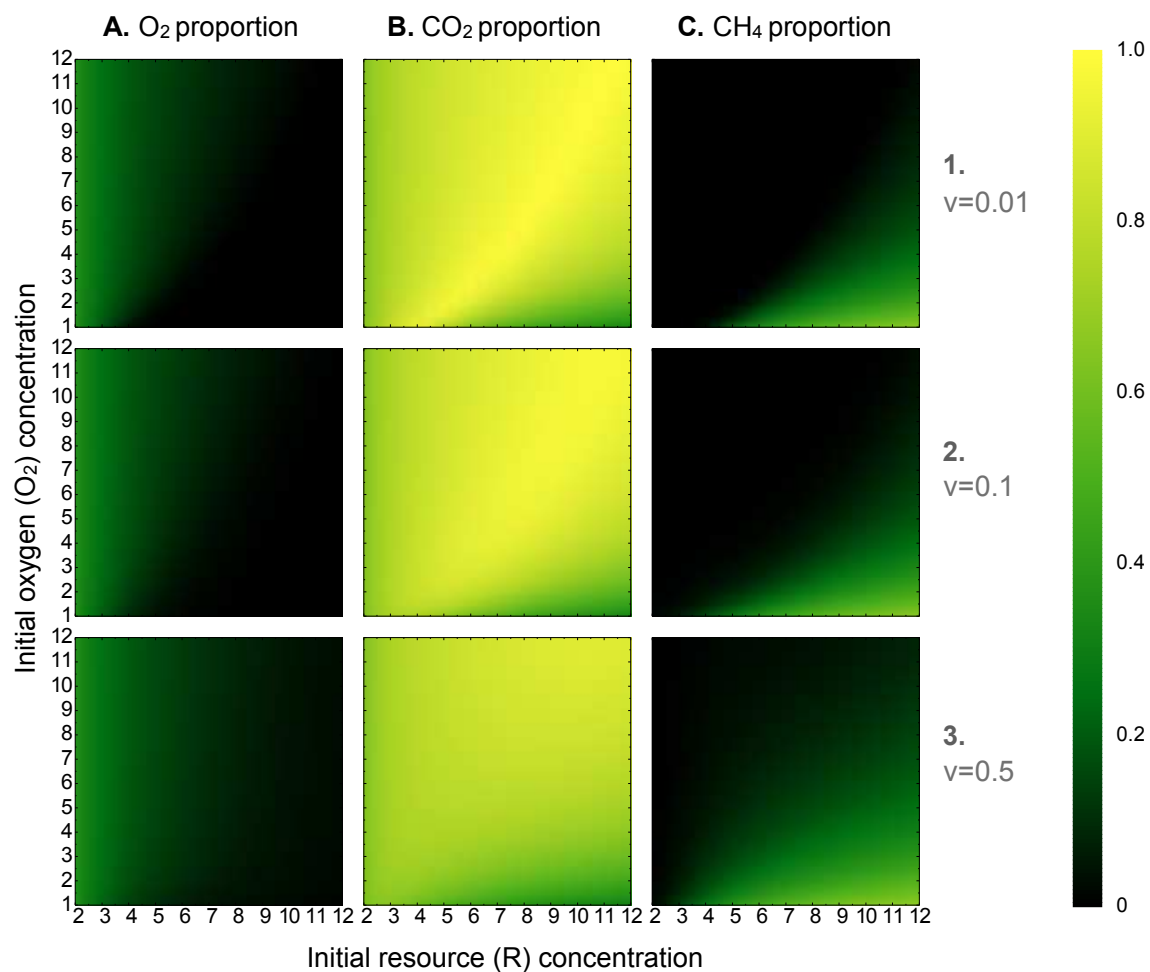


Figure S3.4I: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with energy requirements of aerobic methanotrophs (v) set at 1, 10, and 50% in rows 1-3, respectively.

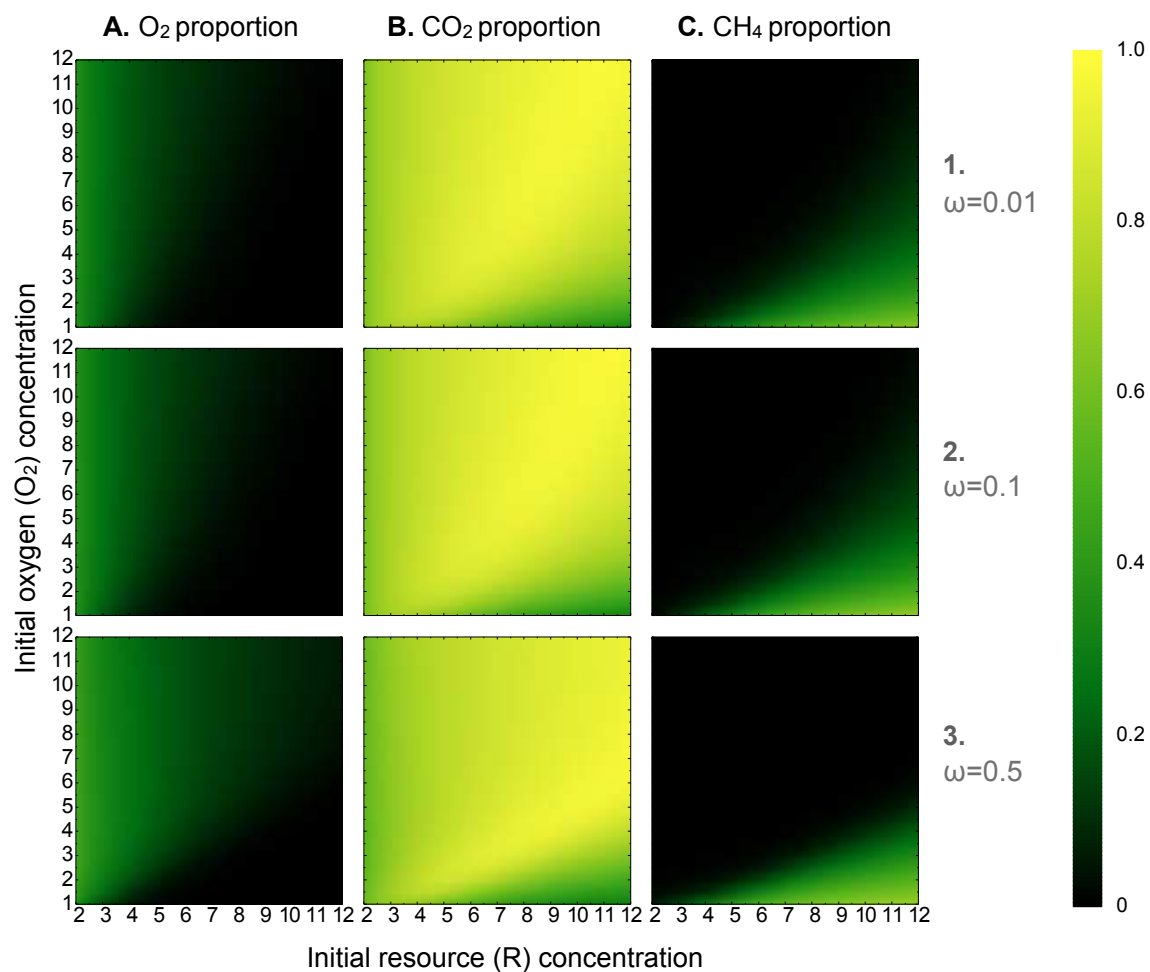


Figure S3.4J: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource utilization efficiency in fermenters (ω) set at 1, 10, and 50% in rows 1-3, respectively.

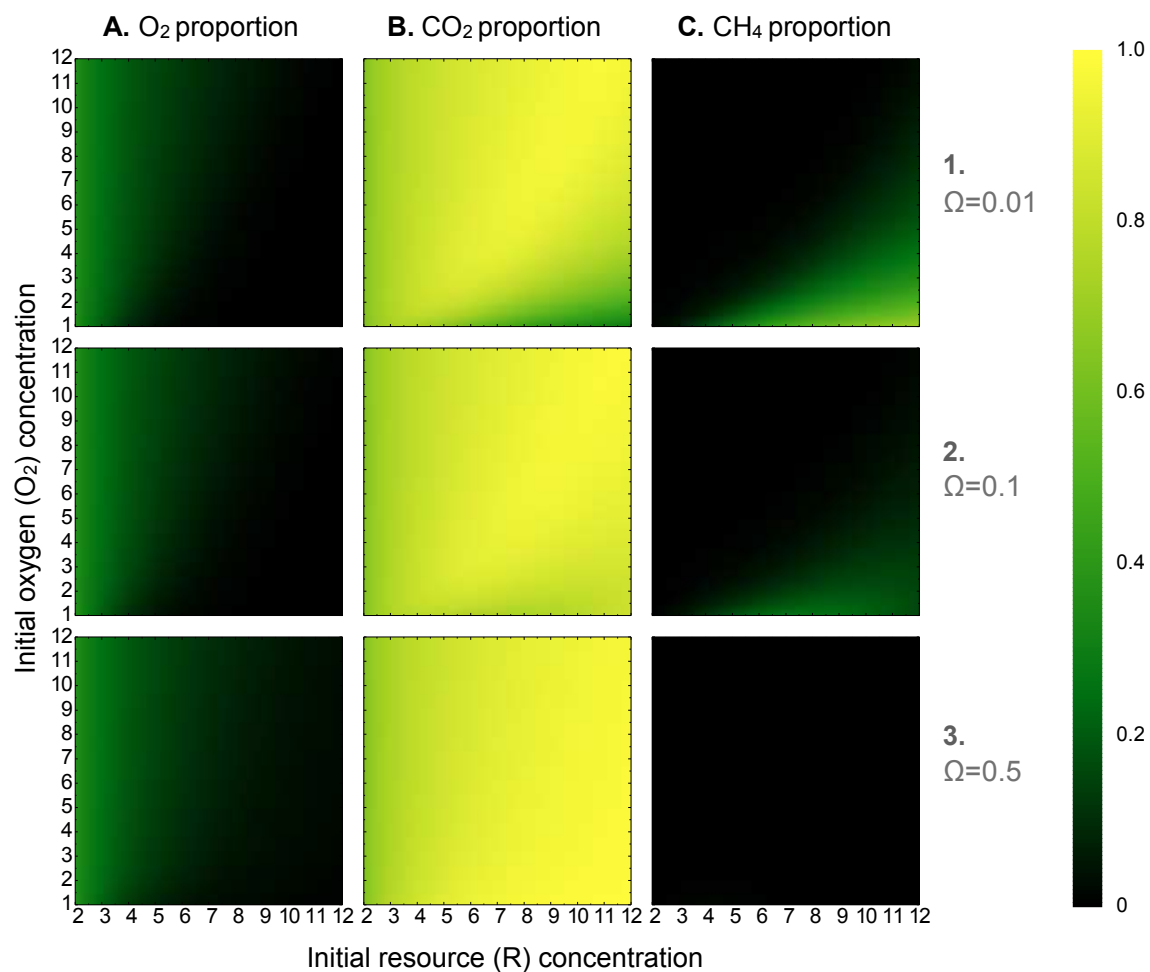


Figure S3.4K: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource utilization efficiency in anaerobic methanotrophs (Ω) set at 1, 10, and 50% in rows 1-3, respectively.

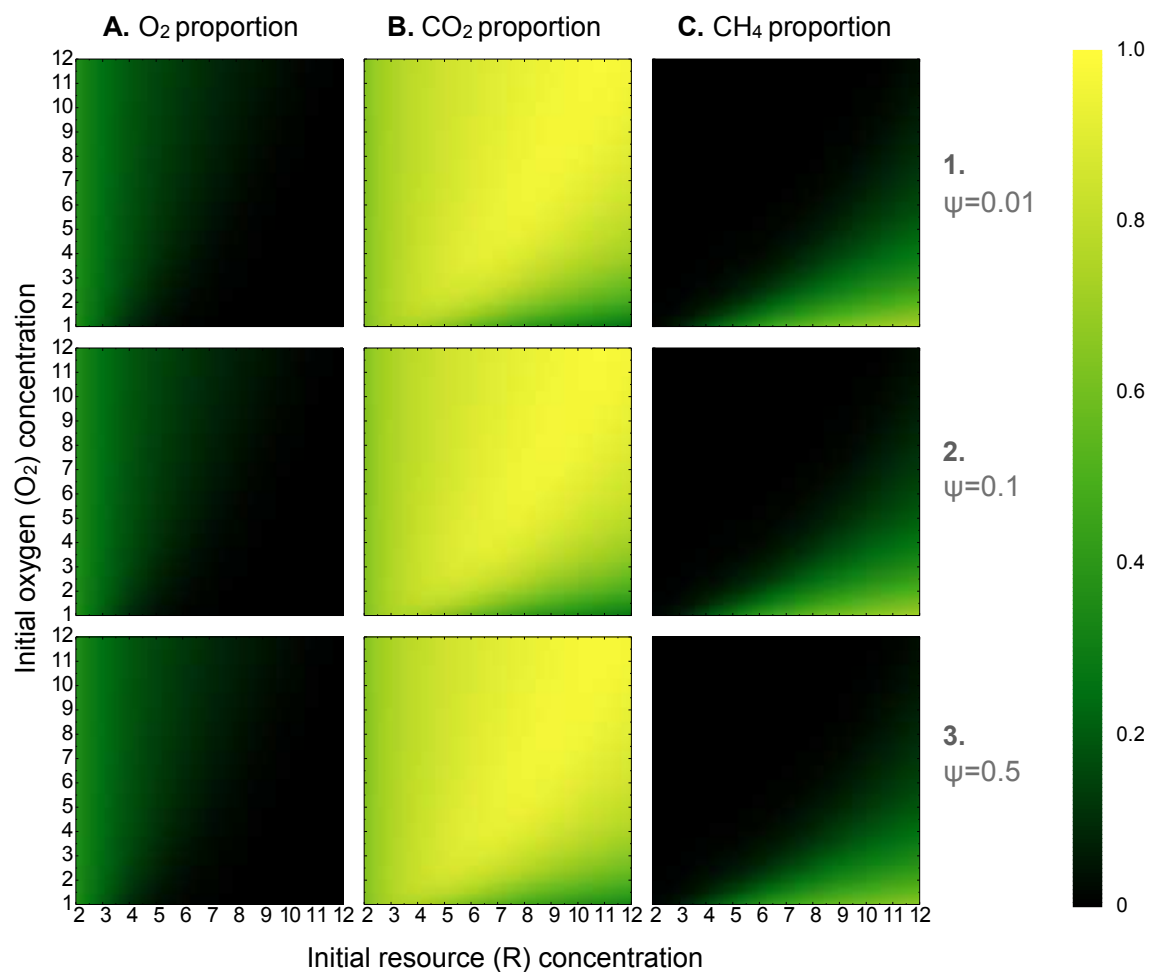


Figure S3.4L: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource incorporation efficiency in anaerobic methanotrophs (ψ) set at 1, 10, and 50% in rows 1-3, respectively.

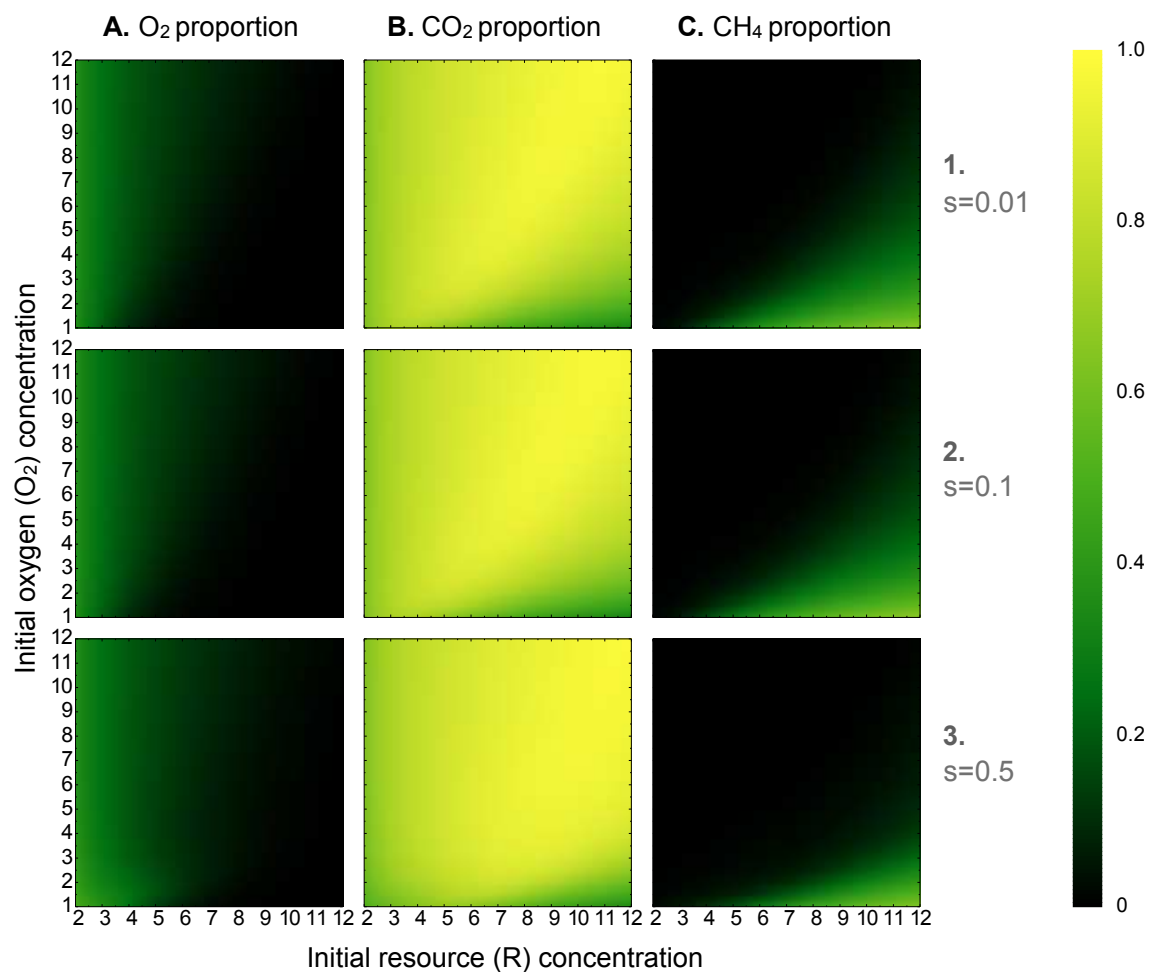


Figure S3.4M: Variations in O₂, CO₂, and CH₄ proportions (in columns A-C) run to steady state at set O₂ and R input values, with energy requirement for aerobic heterotrophs (s) set at 1, 10, and 50% in rows 1-3, respectively.

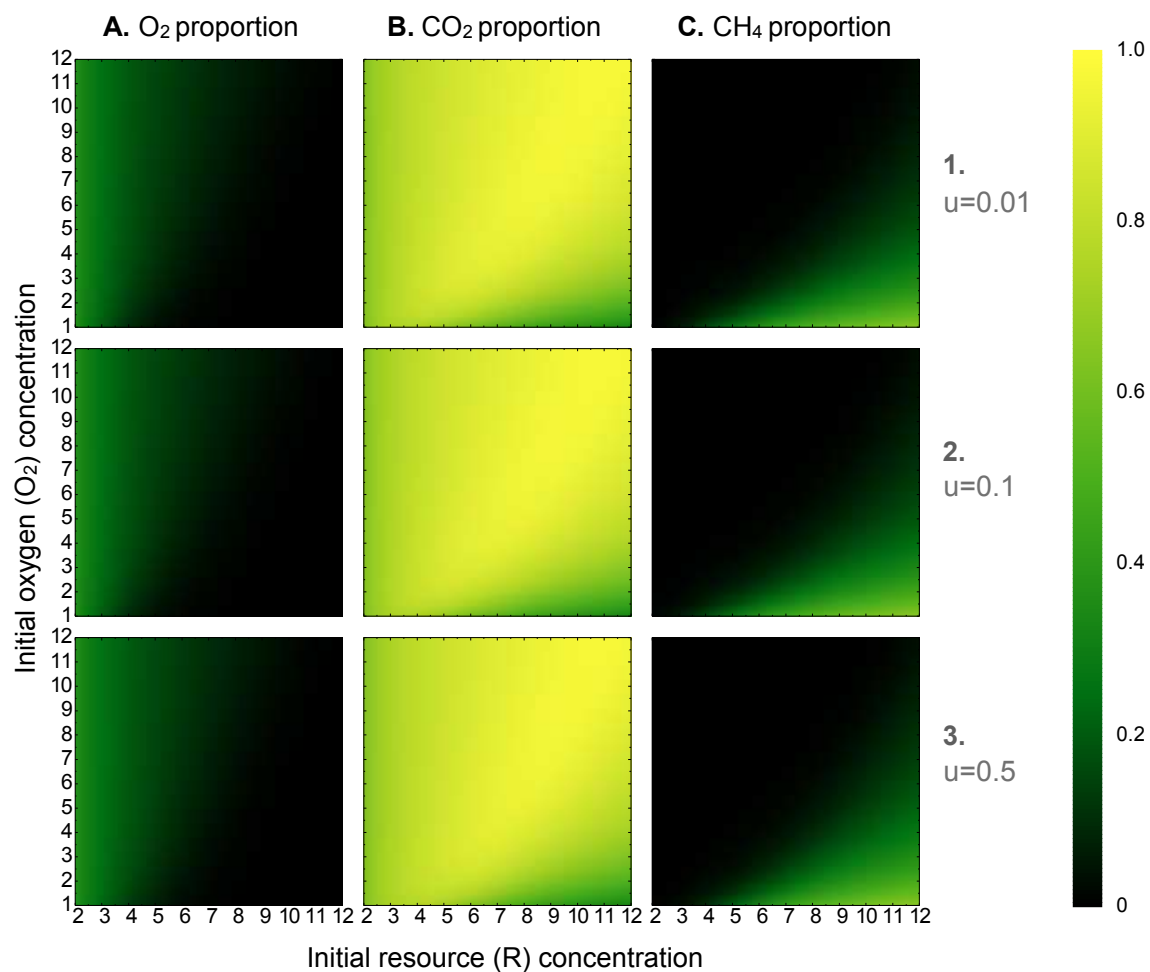


Figure S3.4N: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with energy requirement for methanogens (u) set at 1, 10, and 50% in rows 1-3, respectively.

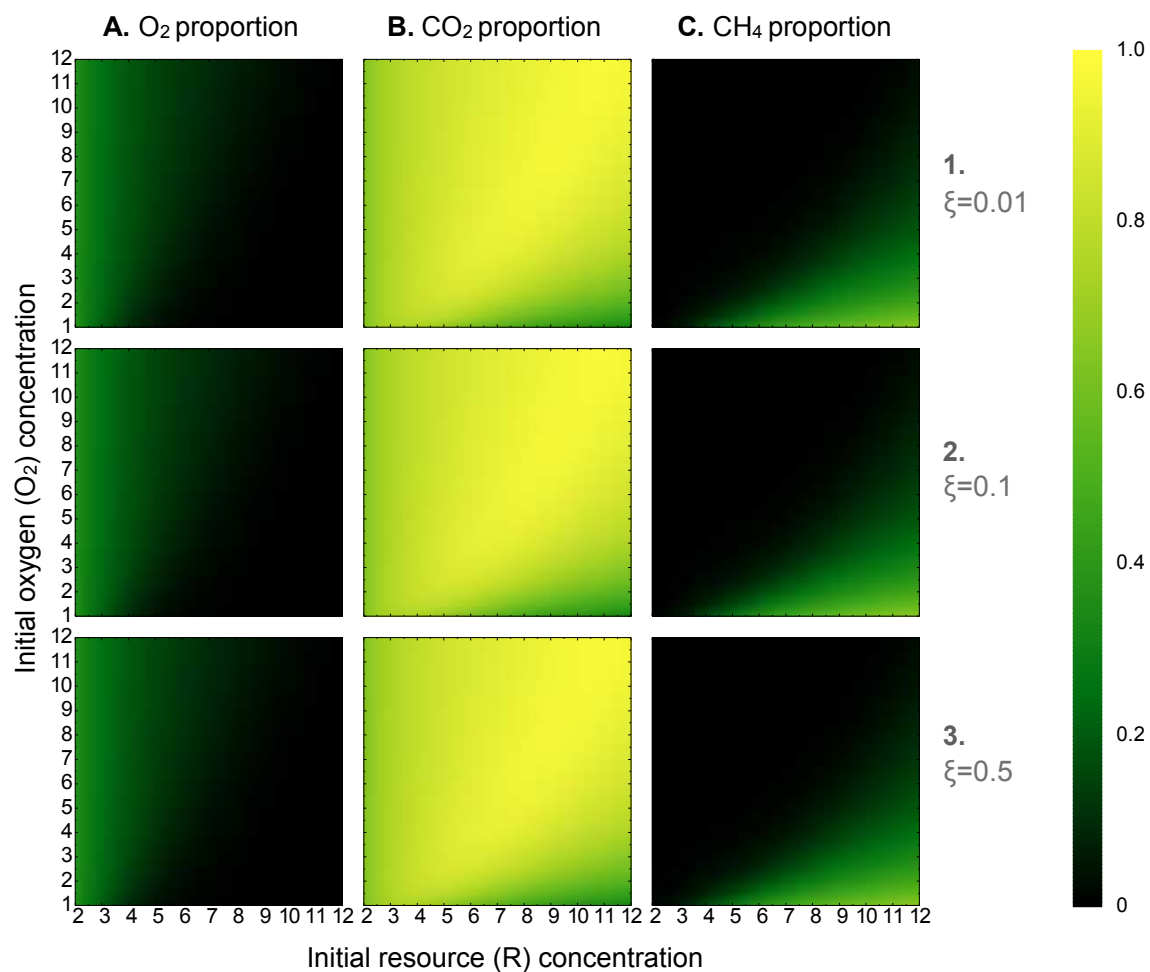


Figure S3.40: Variations in O₂, CO₂, and CH₄ proportions (in columns A-C) run to steady state at set O₂ and R input values, with resource incorporation efficiency for fermenters (ξ) set at 1, 10, and 50% in rows 1-3, respectively.

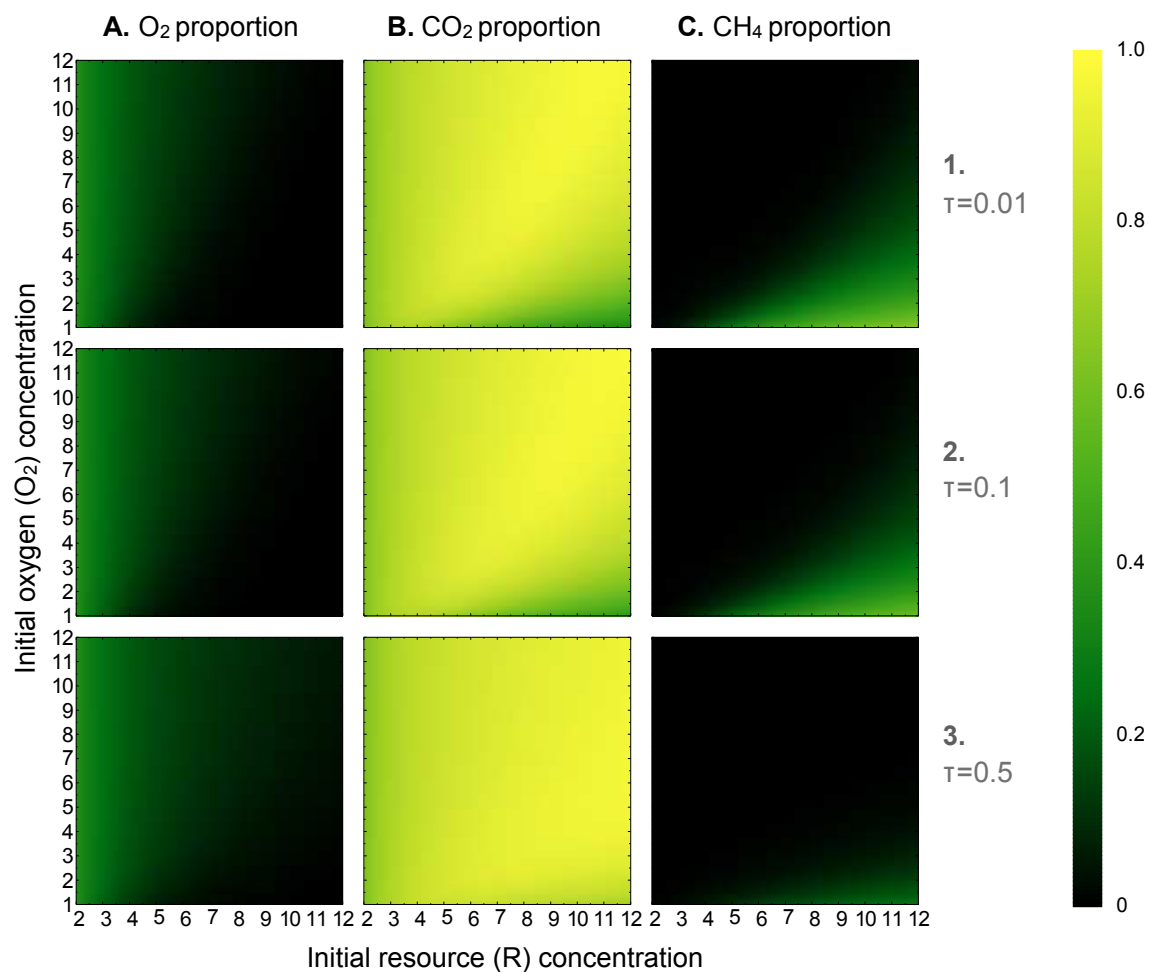


Figure S3.4P: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource utilization efficiency for fermenters (τ) set at 1, 10, and 50% in rows 1-3, respectively.

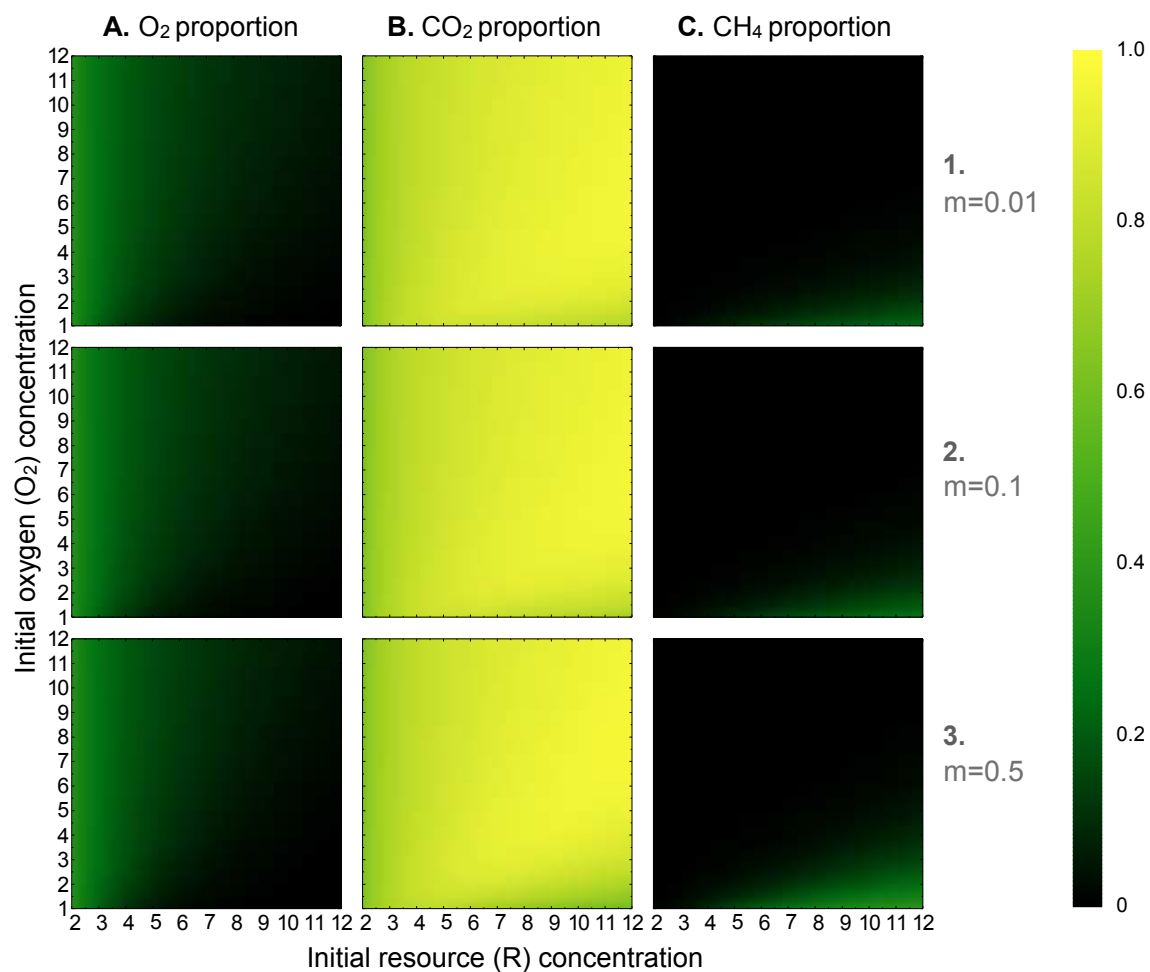


Figure S3.4Q: Variations in O₂, CO₂, and CH₄ proportions (in columns A-C) run to steady state at set O₂ and R input values, with energy requirement for anaerobic heterotrophs (m) set at 1, 10, and 50% in rows 1-3, respectively.

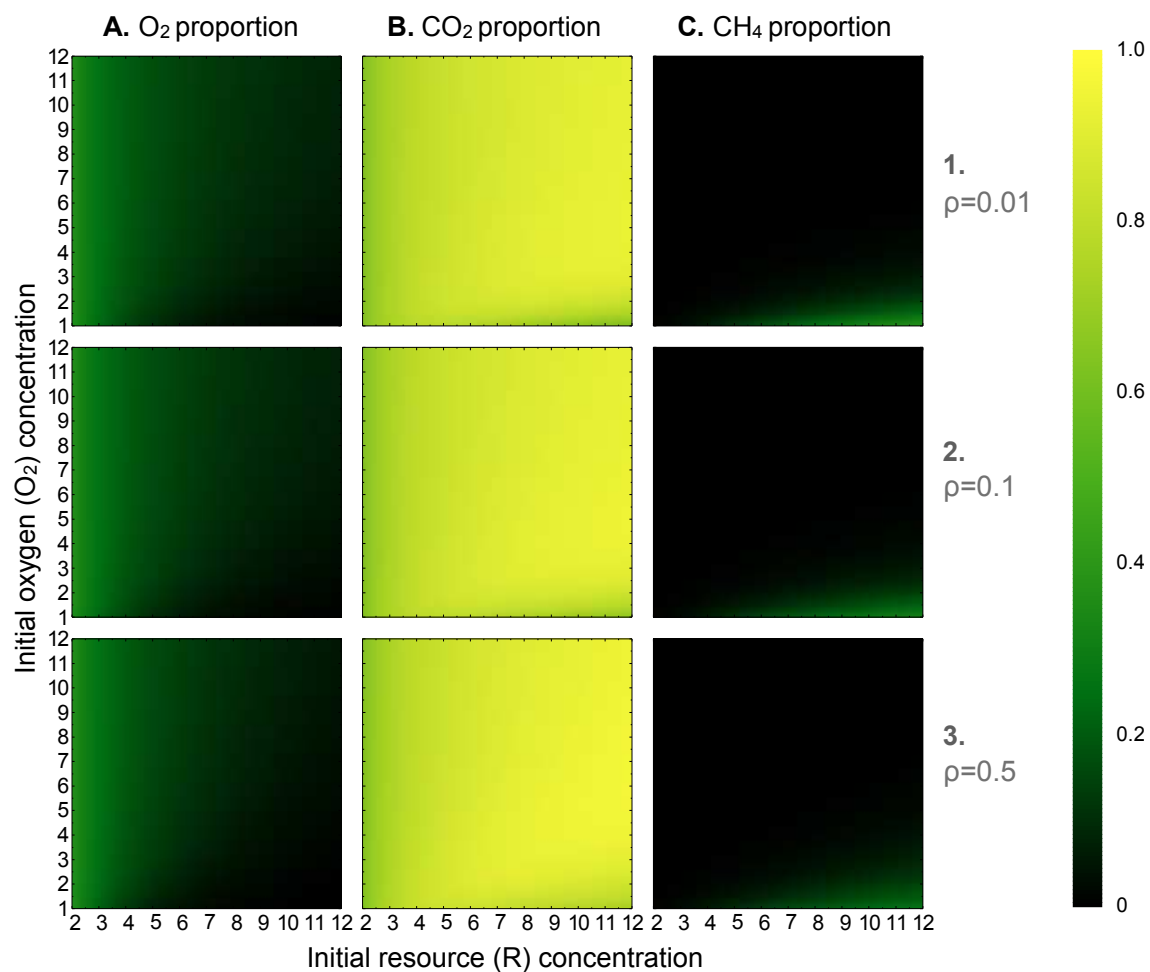


Figure S3.4R: Variations in O_2 , CO_2 , and CH_4 proportions (in columns A-C) run to steady state at set O_2 and R input values, with resource incorporation efficiency for anaerobic heterotrophs (ρ) set at 1, 10, and 50% in rows 1-3, respectively.

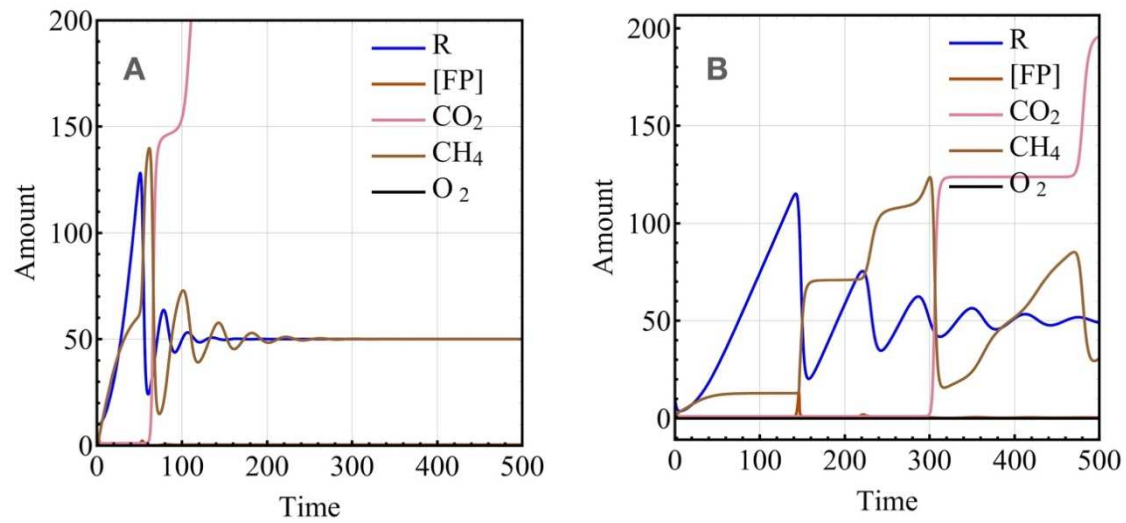


Figure S3.5: Long term dynamic variations in O₂, FP, and CH₄ run at two different initial O₂ concentrations, but constant input of R (resource) - (A) 5 units, (B) 1 unit. The initial O₂ concentrations affect the transient dynamics, but nevertheless the system converges towards the same stable state. These dynamics are reminiscent of the complex cyclic dynamics that we observe in the constant input of both resource and oxygen.

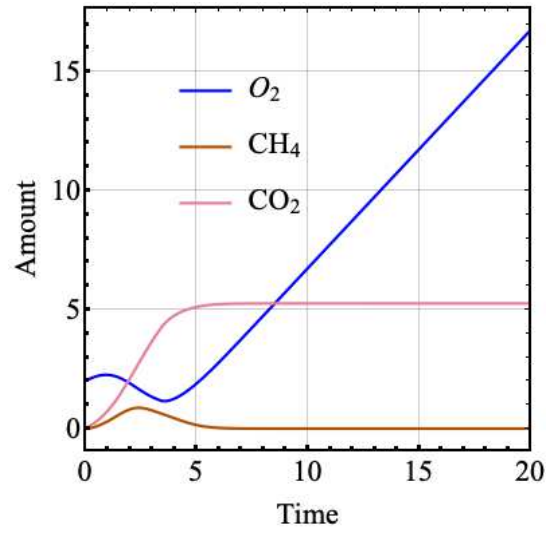


Figure S3.6: Dynamic variations in O_2 , CO_2 , and CH_4 under constant addition of O_2 and fixed initial amount of R (resource). In contrast to figure S5, due to lack of resources, different functional groups crash and the concentrations of gases/substrates (except oxygen) stabilize. Because oxygen is being pumped in continually, it keeps increasing.

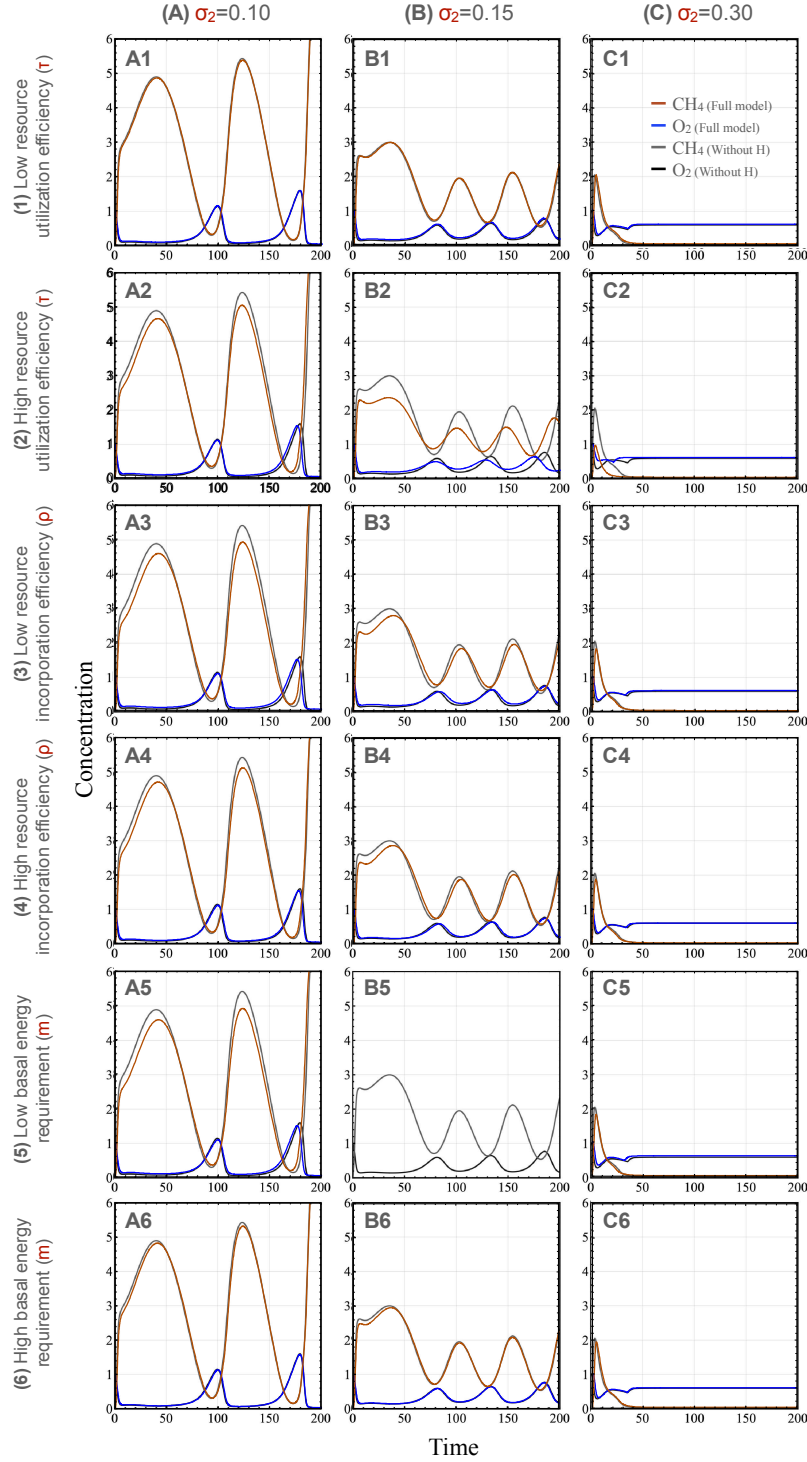


Figure S3.7: A comparison of concentrations of O_2 and CH_4 in time under three different regimes of addition of O_2 (σ_2) (and fixed rate of R (resource) addition (σ_1)) in the reduced model (full model without H) compared with the full model. The dynamics of methane and oxygen are similar in both the models and are mostly controlled by the relative addition rates of oxygen in both the models. This signifies the qualitative robustness of the model to addition or removal of new microbial functional groups (with respect to general behavior)

Chapter 4 : Higher-order effects, continuous species interactions, and trait evolution shape microbial spatial dynamics

Please note that the contents of this chapter have been published as:

Swain*, A., Fussell*, L. and Fagan, W.F., 2022. Higher-order effects, continuous species interactions, and trait evolution shape microbial spatial dynamics. *Proceedings of the National Academy of Sciences*, 119(1).

DOI: <https://doi.org/10.1073/pnas.2020956119>

(*equal contributions; first co-authors)

Abstract

The assembly and maintenance of microbial diversity in natural communities, despite the abundance of toxin-based antagonistic interactions, presents major challenges for biological understanding. A common framework for investigating such antagonistic interactions involve cyclic dominance games with pairwise interactions. The incorporation of higher-order interactions in such models permits increased levels of microbial diversity, especially in communities where antibiotic producing, sensitive, and resistant strains co-exist. However, most such models involve a small number of discrete species, assume a notion of pure cyclic dominance, and focus on low mutation rate regimes, none of which well represents the highly interlinked, quickly evolving, and continuous nature of microbial phenotypic space. Here, we present an alternative vision of spatial dynamics for microbial communities based on antagonistic interactions—one in which a large number of species interact in continuous phenotypic space, are capable of rapid mutation, and engage in both direct and higher-order interactions mediated by production of and resistance to antibiotics. Focusing on toxin production, vulnerability, and inhibition among species, we observe

highly divergent patterns of diversity and spatial community dynamics. We find that species interaction constraints (rather than mobility) best predict spatiotemporal disturbance regimes, whereas community formation time, mobility, and mutation size best explain patterns of diversity. We also report an intriguing relationship among community formation time, spatial disturbance regimes, and diversity dynamics. This relationship, which suggests that both higher-order interactions and rapid evolution are critical for the origin and maintenance of microbial diversity, has broad-ranging links to the maintenance of diversity in other systems.

Introduction

Understanding the origin and maintenance of species diversity is a long-standing biological question, and understanding diversity through the lens of species interactions has been deemed key (Hutchinson, 1961; May, 1972; Chesson, 2000; McCann, 2000; Shores et al., 2008). Natural communities possess high species diversity, often much higher than expected from a 'simple' Darwinian interpretation of species interactions (Reichenbach et al., 2007). Extraordinary levels of diversity exist in several macroscopic systems (e.g., coral reefs, tropical forests), but remarkable levels of diversity also exist in microscopic systems, among microbial communities. High levels of diversity in microbial systems are especially intriguing given the ubiquity of antagonistic interactions among microbes, such as the production of antibiotics/toxins against one another (Cordero et al., 2012; Ghoul and Mitri, 2016; García-Bayona and Comstock, 2018; Hibbing et al., 2010; Peterson et al., 2020).

Theoretical and empirical studies have explored microbial interactions with the goal of understanding the mechanisms of species coexistence in hyper-diverse microbial systems. Major examples include models with: **(A)** mutualistic and cross-feeding interactions (Hoek and Merks, 2017); **(B)** resource-based competitive interactions (Butler and O'Dwyer, 2018); **(C)** a

combination of mutualistic and competitive interactions (Butler and O'Dwyer, 2018, 2020; Coyet et al., 2015; Momeni et al., 2017); **(D)** stochastic spatio-temporal processes (Chase, 2003; Jeraldo et al., 2012); and **(E)** antagonistic interactions due to toxin production (Czárán et al., 2002; Kelsic et al., 2015; Reichenbach et al., 2007).

Previous studies focusing on the mutualistic, cross-feeding and resource-based competitive interactions (points **A-C** above) have demonstrated that positive interactions arising from resource exchange can affect key criteria governing the relationship between community diversity and stability. Examples of such criteria are that species richness is inversely related to the strength of random pairwise interactions in stable communities (May, 1972) and communities constructed with random pairwise interactions feature an upper bound on diversity (Allesina and Tang, 2015; Gross et al., 2014; Sinha and Sinha, 2005). Additional key findings from the resource-based framework are that cooperative (metabolic) interactions can arise spontaneously in complex microbial communities (Libby et al., 2019) and that a cooperative versus competitive dichotomy of microbial communities may exist, depending on conditions (Machado et al., 2021). Sometimes, the interactions arising from resource exchange allow systems to bypass classic stability-diversity relationships (Butler and O'Dwyer, 2018, 2020). Collectively, these results have tremendously improved our understanding of the factors influencing the development and persistence of microbial community diversity. However, resource-based perspectives fail to account for a major class of well-known antagonistic interactions that exist among microbes, mediated through bacterial toxins or antibiotics (also termed as bacteriocins and related compounds). Toxin-based interactions are ubiquitous in both natural microbial communities and controlled settings (Cordero et al., 2012; García-Bayona and Comstock, 2018; Ghoul and Mitri, 2016; Hibbing et al., 2010),

are found in all known phyla of bacteria, come in diverse forms, and play a central, yet usually underappreciated role in shaping microbial community structure (Peterson et al., 2020).

Here, we focus on modeling microbial communities structured by antagonistic interactions among antibiotic/toxin producing, sensitive and resistant species. Prior research in this area has relied on the framework of cyclic dominance (also termed non-transitive systems or rock-paper-scissors systems) and related techniques in evolutionary game theory (Czárán et al., 2002; Kelsic et al., 2015; Frey, 2010; Hibbing et al., 2010; Kerr et al., 2002; Szolnoki et al., 2014). Early models using cyclic dominance explored community stabilization, increased co-existence, and maintenance of diversity in spatially structured environments (Czárán et al., 2002; Durrett and Levin, 1998; Kerr et al., 2002; Kirkup and Riley, 2004). However, because these models focused on pairwise interactions, they generally did not account for other kinds of interactions that occur when intermixed species coexist at very small spatial scales (García-Bayona and Comstock, 2018; Ghoul and Mitri, 2016; Green et al., 2006; Raynaud and Nunan, 2014; Reichenbach et al., 2007; Vetsigian et al., 2011). For example, real microbial species are routinely involved in multi-species interaction systems including so-called ‘higher-order interactions’ in which the interactions between two species can be modulated by other species (Bairey et al., 2016; Grilli et al., 2017). One simple example of such a behavior occurs when an antibiotic produced by one species, that inhibits the growth of a competing (sensitive) species, can be attenuated by a third species that can degrade the antibiotic (Abrudan et al., 2015; Kelsic et al., 2015; Perlin et al., 2009). Here, the third species modulates the interaction between the antibiotic-producing and antibiotic-sensitive species, without impacting either of them directly (Bairey et al., 2016), thus creating a higher-order interaction.

Models using cyclic dominance with higher-order interactions for a small number of static (i.e., non-evolving) species have demonstrated increased stability and diversity in both spatially structured and well-mixed communities (Czárán et al., 2002; Durrett and Levin, 1998; Kelsic et al., 2015; Kerr et al., 2002; Kotil and Vetsigian, 2018; Reichenbach et al., 2007), and findings from such models have been verified empirically (Kelsic et al., 2015). In contrast, horizontal gene transfer and mutation blur the boundaries among microbial strains and species, reflecting the continuous nature of phenotypes in microbial assemblages. Scaling up analytic models of cyclic dominance to species-rich scenarios is not possible because the incorporation of more than a few discrete species makes the dynamics extremely complex and hard to interpret using equation-based approaches (Kelsic et al., 2015).

Despite its prominence in modeling studies, pure cyclic dominance, in which there is a ‘closed’ loop of dominance, is rare in nature (Friedman et al., 2017; Higgins et al., 2017; Park et al., 2020; Wright et al., 2016). In contrast, rapid evolution routinely leads to model systems featuring semi-cyclic or non-cyclic patterns in which higher-order interactions, such those involving antibiotic producing, sensitive, and resistant species, are common (Grilli et al., 2017; Park et al., 2020). Therefore, the notion of pure cyclic dominance used in earlier studies of antagonistic microbial dynamics needs to be supplemented with evolving, ‘mixed’ patterns of dominance, in order to better model microbial community interactions.

Prior studies of community-level eco-evolutionary dynamics, which have explored the evolution of species interactions abstractly or the specific traits that determine such interactions (Ginzburg et al., 1988; Guill and Drossel, 2008; Loeuille and Loreau, 2005), focused on niche-based food web models (i.e., those involving resource consumption) rather than the non-resource based antagonistic interactions that characterize our work. Even when studies did investigate

antagonistic interactions and resultant species dynamics, they emphasized very slow mutation regimes and assumed a separation of ecological and evolutionary timescales (Diekmann, 2002; Gorter et al., 2020; Tarnita, 2018). Such assumptions do not work very well for microbial systems, where high mutation rates and extensive horizontal gene transfers are commonplace, creating an overlap of ecological and evolutionary timescales (Gorter et al., 2020; Hall et al., 2017; Le Roux and Blokesch, 2018; Tarnita, 2018). These eco-evolutionary feedbacks can affect the nature of trait evolution (48) and even destabilize species interactions (Rodríguez-Verdugo and Ackermann, 2021). Recent models exploring these fast-evolving regimes in antibiotic-mediated antagonistic communities have uncovered mechanisms responsible for *de novo* assembly of diverse microbial communities with higher-order interactions, but such results are qualitatively different from what is possible when mutation is rare (Kotil and Vetsigian, 2018; Tarnita, 2018). What is needed, then, are studies of community assembly in models where a large number of species interact in continuous phenotypic space, are capable of rapid mutation, and engage in both direct and higher-order interactions.

Here, we integrate three major themes concerning the dynamics of microbial communities structured by antagonistic interactions, specifically higher-order interactions, a continuous view of microbial trait space, and an eco-evolutionary perspective on community assembly wherein evolution can happen rapidly. This integrated perspective allows us to explore mixed, rather than strict, patterns of dominance among a large number of biologically realistic species (hundreds to thousands; constrained only by computational power) to study microbial communities of toxin/antibiotic producing, sensitive, and resistant species. Within an agent-based modeling framework, we use a continuous species parametrization model with a wide range of mutational sizes to explore spatio-temporal community assembly, diversity, and stability (see **Figure 4.1** and

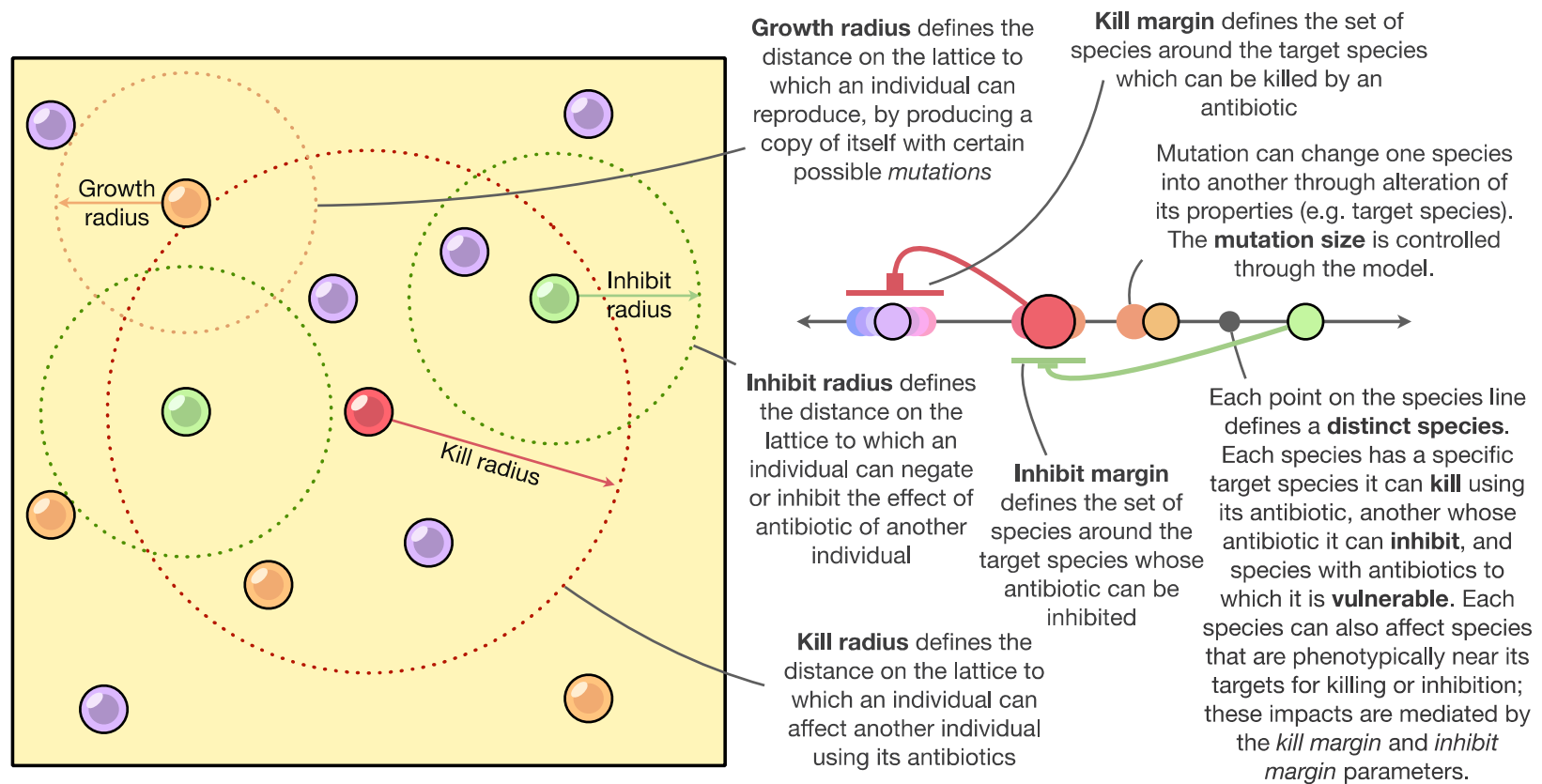


Figure 4.1: A conceptual representation of the model, with descriptions of all the parameters involved. See Methods for details about implementation.

methods). Parameters are defined at a global level and are divided into two categories: spatial (which define rules and properties of spatial interactions) and species-level (which define the toxin interactions in the species space). The continuous (phenotypic) species space generalizes intra-species relationships of toxin production, vulnerability, and resistance that arise randomly in populations through mutation and interspecies relationships.

We find that species-level properties, such as constraints on higher-order interactions among species, affect the spatial structure of microbial dynamics more than spatial factors like mobility and the effective distance over which individuals kill or inhibit each other. In contrast, nonspatial metrics of community diversity reflect complex interactions among species-level and spatial parameters, especially mobility and mutation size, plus community formation time (the time it takes to form a stable community; see Kotil and Vetsigian, 2018). Our attention to community formation time, which provides an opportunity for characterizing transient dynamics rather than just focusing on stable states or the long-term maintenance of diversity (Coyte et al., 2015; Friedman et al., 2017; Tarnita, 2018), is intentional. It turns out that community formation time is a very good predictor of diversity dynamics, but this time duration cannot be estimated well from other parameters, pointing at emergent complex properties of community assembly.

Results

After running 10.49 million simulations across a broad range of parameter combinations, we observed strong differences in spatial heterogeneity among runs. We classified this behavior into three major categories depending only on the observed level of spatial disturbance (low, medium, and high disturbance; **Figure 4.2**, Methods). These three categories featured both between- and within-category differences in community spatio-temporal dynamics, but exhibited similar

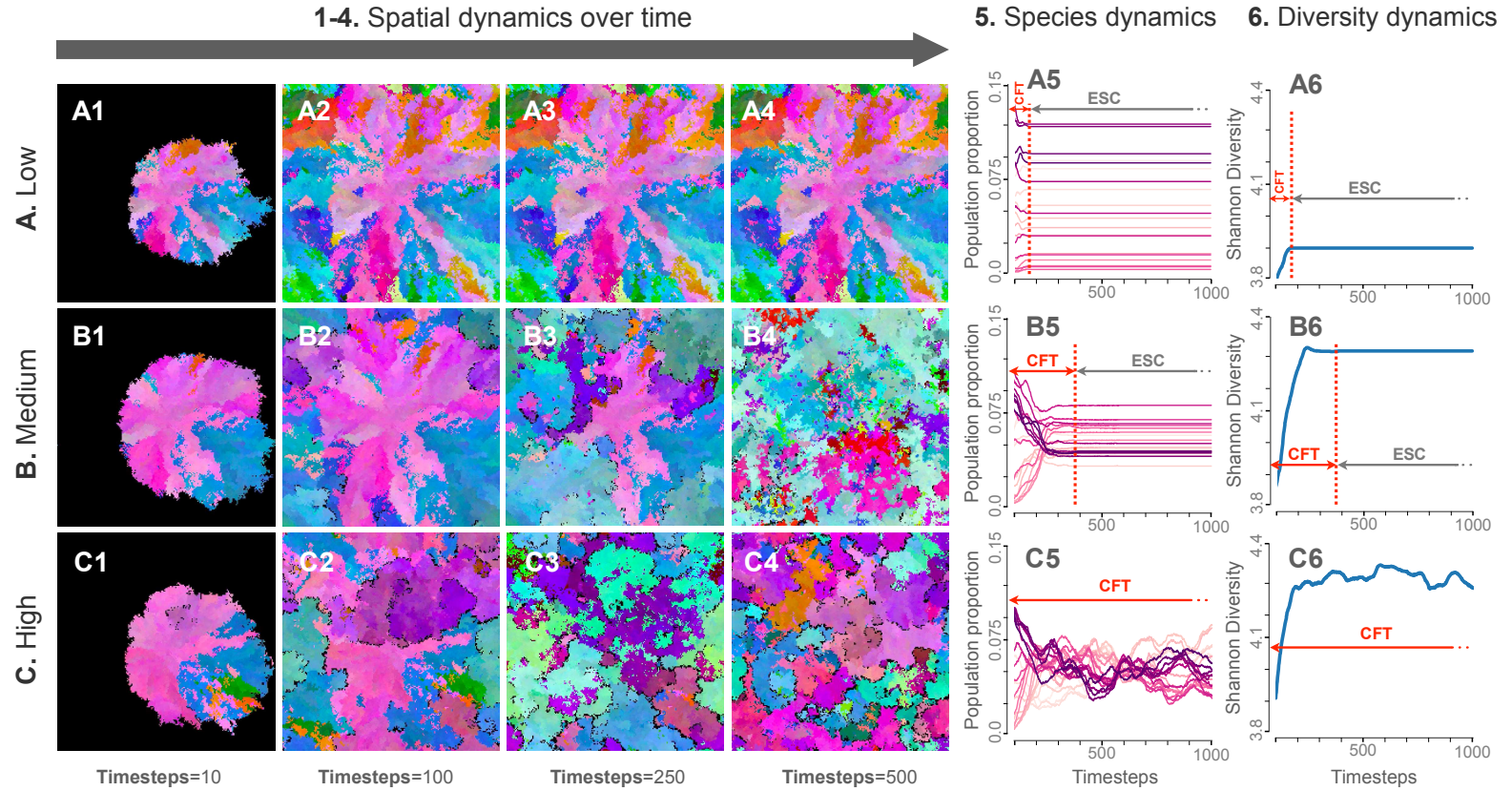


Figure 4.2: System dynamics in three example conditions of initial conditions pertaining to different regimes of spatial disturbance behavior (Low, Medium and High) (rows A-C). Snapshots of each simulation at different timesteps (10, 100, 250 and 500) are shown in columns 1 through 4. The respective species dynamics (where different colors represent different species ‘bins’, see Methods) and diversity dynamics are represented in columns 5 and 6, respectively. In columns 5 and 6, the time taken for the simulation to stabilize (i.e., the community formation time or CFT) is denoted with a vertical red dotted line. The community existing after the system has reached the CFT is termed an eco-evolutionary stable community (ESC).

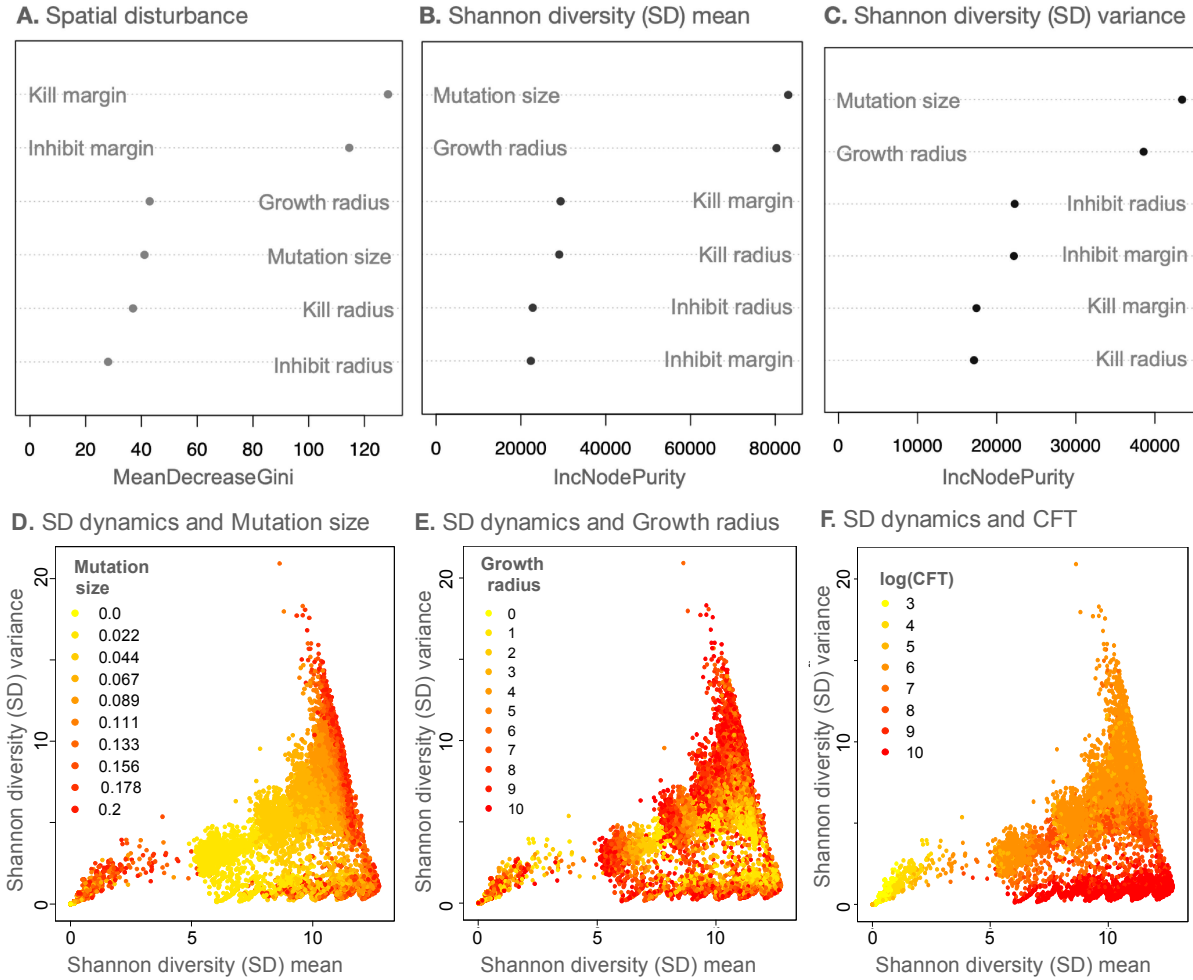


Figure 4.3: Dependence of model dynamics on different parameters and community formation time (CFT). (A) shows the variable importance score for all the parameters from a random forest classification (with 10,000 trees) on the three categories of spatial disturbance (OOB=10.2%). (B) and (C) denote the variable importance scores from random forest regression models (with 10,000 trees) for Shannon diversity (SD) mean and variance respectively. (D)-(E) show the SD mean-variance plots overlaid with mutation size, growth radius, and community formation time (CFT) values of the respective simulations. For (A)-(C), IncNodePurity (used in regression random forests) and MeanDecreaseGini (used in classification random forests) refer to how well a parameter explains the prediction variable in question.

disturbance regimes within categories (**Figure 4.2**). The continuous trait axis and the mutable nature of traits together allow for the simulation and visualization of hyper-diverse communities, contrasting with the small, fixed number of discrete, immutable species in previous works (see Kelsic et al., 2015; Frey, 2010; Reichenbach et al., 2007).

The time that each simulation took to stabilize (i.e., to form a stable community) is termed the community formation time (CFT; see Methods and Kotil and Vetsigian, 2018), and the community that thus assembled is called an eco-evolutionary stable community (ESC). The incorporation of mutations allows for an exploration of community assembly, transient dynamics (such as CFT and its relation to community properties over time), and complex spatial patterns that was not possible using previous frameworks (**Figure 4.2**) (Kelsic et al., 2015; Kotil and Vetsigian, 2018).

To understand how parameters relate to the spatial disturbance categories, we performed a random forest classification using the six model parameters (*kill radius*, *inhibit radius*, *growth radius*, *kill margin*, *inhibit margin* and *mutation size*). The model had an out-of-bag error of 10.2% and identified *kill margin* and *inhibit margin* as the best predictors of the disturbance classification (**Figure 4.3A**). Note that *kill margin* and *inhibit margin* are global species-specific parameters.

The next most important parameter was *growth radius*, which is a proxy for mobility in our model (**Figure 4.1**), followed by *mutation size*, *kill radius*, and *inhibit radius*. These findings showed that, except for *growth radius*, other spatial parameters played a minor role in determining the structure of spatial disturbance in the model systems.

To assess non-spatial metrics of community assembly, we performed a random forest regression on the Shannon diversity (SD) mean (i.e., mean Shannon diversity (SD) for the period from timestep=0 to the CFT) and the associated SD variance using the same six simulation parameters (and an associated species bin size of 0.05 for discretizing species; for more details on

species binning, see Methods). That random forest model explained 68.4% of the variance for the SD mean and 70.2% for the SD variance. *Mutation size* and *growth radius* were the two best predictors of both outcome variables (**Figure 4.3B,C**). Among the less important parameters that affect SD mean and variance, we note that *kill margin* and *kill radius* play a larger role in determining SD mean whereas inhibition processes better predict SD variance (**Figures 4.3B,C, S4.1**).

From a naïve interpretation, one would expect *mutation size* to be a good predictor of increasing diversity, as a higher *mutation size* allows for the possibility of greater variation in microbial phenotypic space and therefore the binned SD would be expected to be higher. However, after overlaying the plot of SD mean versus variance with the respective *mutation size* values, the picture is largely muddled (**Figure 4.3D**). We saw even more mixed results after overlaying the growth radius values on the SD mean-variance plot (**Figure 4.3E**). Overlaying other less important model parameters on the SD mean-variance plot showed no straightforward pattern at all (**Figure S4.2**).

Interestingly, upon overlaying the SD mean-variance plot with the values of the respective CFTs, a stronger pattern emerged. Note the clustering by color gradient of the points due to CFT in SD mean-variance space (**Figure 4.3F**). Although CFT is not a parameter, it is an emergent outcome of the community assembly process and exploring it further provided insights into the diversity formation process itself.

To formally detect clusters in the SD mean-variance plot with respect to CFT, we used k-means clustering optimized for the number of clusters using multiple methods (Methods) and identified four clusters (**Figures 4.4A, S4.3**). These clusters pertain to (1) Low SD mean and SD variance, (2) High SD mean and SD variance, (3) High SD mean and medium SD variance and, (4) High SD mean and low SD variance, in order of increasing values of CFT.

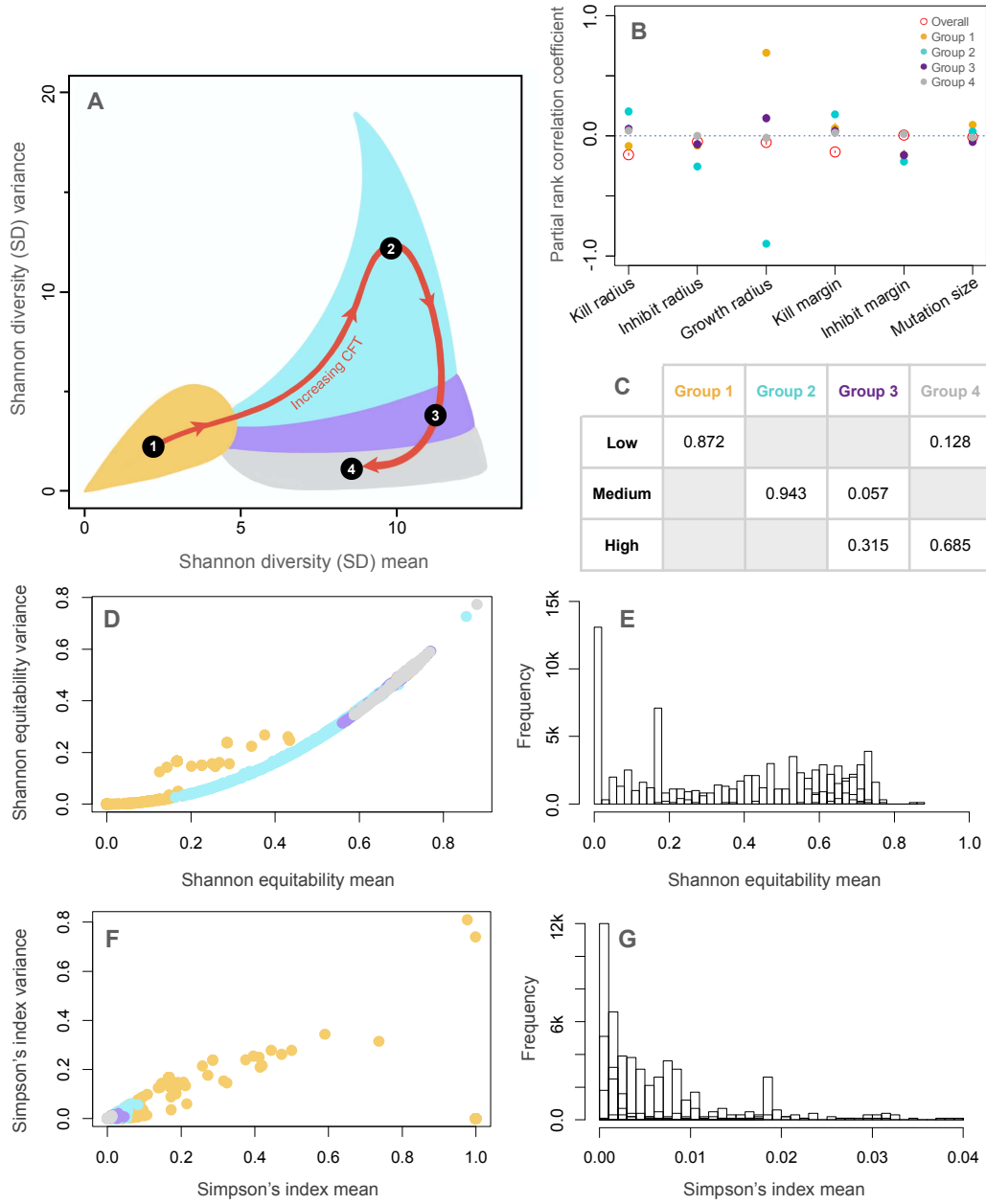


Figure 4.4: Effects of Community formation time (CFT) on the spatial diversity dynamics. **(A)** depicts the four optimal clusters (groups) of SD mean-variance space found using k-means clustering (see Figure S3). **(B)** shows the results from the partial rank correlation coefficient (PRCC) analyses of the four groups and the overall data. **(C)** shows the proportion of spatial disturbance regimes present in different groups. **(D)** and **(F)** show the distribution of the values of mean and variance of Shannon Equitability and Simpson's index, respectively, for a smaller subsample of the runs, with points colored by groups from **(A)**. **(E)** depicts the group-wise histogram of Shannon equitability and **(G)** the histogram of Simpson's index (for clarity, only the range between 0 and 0.04 is shown).

On performing random forest regression predicting CFT using the six-parameters for the whole dataset, we found a low explanation of variance (22.3%), with the *kill parameters* (*kill radius*, and *kill margin*) being the strongest predictors. However, upon performing the analysis separately on the four CFT groups, a stronger pattern emerged (**Figure S4.4**). For group 1, the regression explained 30.2% of the variance with *growth radius* being the strongest predictor, and other variables playing a smaller part, but nevertheless the second and third best predictors are *kill parameters* (**Figure S4.3A**). Group 2 regression explained 90.6% of the variance with again *growth radius* as the strongest predictor with other parameters making almost no contributions (**Figure S4.4B**). For group 3 and 4, the regression explains 27.2% and 6.0% of the variance respectively, with *growth radius* again having most impact on prediction (**Figure S4.3C,D**). This pattern of results is unusual because the overall regression did not show any strong dependence on *growth radius* (**Figure S4.3E**).

To explore this dilemma further, we performed a partial rank correlation coefficient analysis on this data pertaining to CFT (**Figure 4.4B**). We found that the sign and strength of the dependence of *growth radius* for CFT groups vary widely; hence, the random forest regression on the whole dataset misses this signal. *Growth radius* had a strong positive PRCC with CFT in group 1, a very strong negative PRCC with CFT in group 2, a weakly positive PRCC in group 3 and a negligible one in group 4. Other parameters had relatively smaller PRCC values as compared to *growth radius* values for group 1 and 2. *Kill radius* and *kill margin* had overall higher PRCC values than the other parameters and thus were identified by the random forest algorithm (**Figures 4.4B, S4.4E**). Moreover, on calculating PRCC of parameters on SD mean and SD variance, we found that the general values are similar for both SD metrics, but what differentiates SD mean and SD variance values the most is *growth radius* (**Figure S4.1**).

To link the dependence of diversity dynamics on CFT with the spatial disturbance regimes, we compared what proportion of the three spatial disturbance regimes (low, medium, and high) fall into the four CFT-classified groups from SD mean-variance space (**Figure 4.4C**). We found that the low disturbance regime was most associated with group 1 and, to a much lesser extent, group 4. The medium disturbance regime was primarily dominated by group 2, with a small contribution from group 3. Lastly, the high disturbance regime was dominated by groups 3 and 4. These results can be interpreted in the light of Shannon equitability and Simpson's index values for the simulations (**Figures 4.4D-G**). Transitioning from Group 1 to 4 reveals an increase in mean Shannon's equitability (**Figure 4.4D,E**) and a decrease in mean Simpson's index (**Figure 4.4F,G**). One interpretation here is that higher equitability among species (i.e., higher values of Shannon equitability), which can occur when comparatively few species dominate the community (i.e., lower Simpson's index), results in simulations that take longer to equilibrate (i.e., longer CFTs). For example, communities in Group 4 take longer to equilibrate (longer CFTs) as the species in those communities are more nearly equitable, with comparatively fewer species whose interactions exercise control over the community. In contrast, Group 1 has low equitability, implying a higher level of community dominance by some players, and therefore, those simulations settle down quickly (low CFTs). Collectively, these results highlight some important, and novel, connections between temporal spatial patterns and diversity dynamics, especially through the lens of transient phenomena (see Hastings, 2001; Hastings et al., 2018).

All the results in this section were performed using a species bin size (for discretization of species for diversity metrics; see Methods) of 0.05, but we repeat the analyses at three other reasonable bin sizes (0.03, 0.07 and 0.09; see Methods) and find almost no changes in the structure of the results and their interpretations (see **Figure S4.5**) in any of the cases. The space of SD mean

and variance shrinks in scale (**Figure S4.5C**) as the bin sizes increase, which is expected because increasing bin sizes results in fewer distinct species for analysis. However, the categorization of spatial disturbance remains unaltered because binning does not affect the actual dynamics of the model, but only how we quantify those dynamics through discretization. Therefore, the dependence of CFT on parameters remained unaltered (**Figure S4.5D**). The categorization of SD mean and variance into groups, their relation to spatial disturbance regimes, and their dependence on parameters also varied little (**Figures S4.5A,B,E**).

Discussion

A continuous species model with simple rules for interactions based on antibiotic production, inhibition, and vulnerability can produce a wide array of complex dynamics and generate hyper-diverse, persistent communities. Despite enormous heterogeneity in behavior, certain general patterns in the assembly and stability of these hypothetical microbial communities stand out.

The diversity patterns seen across time were expressed in the form of mean Shannon diversity (SD) and its associated variance. The phase space of these two metrics shows a non-linear relationship between the metrics. Random forest regressions revealed that *mutation size* and *growth radius* were the best predictors of mean SD diversity and its variance (**Figures 4.3B,C**). In other words, the mutation size and mobility of species jointly determine how diversity dynamics play out in these theoretical communities.

Mobility controls diversity dynamics in numerous previous studies (Kelsic et al., 2015; Frey, 2010; Reichenbach et al., 2007; Szolnoki et al., 2014). In particular, small amounts of mobility enhance diversity, whereas large amounts of mobility jeopardize it (Reichenbach et al.,

2007; Szolnoki et al., 2014). Reduced mobility results in spatially structured populations where co-existence is easier to maintain, as seen in experimental works on competing *E. coli* strains (Kerr et al., 2002, 2006; Weber et al., 2014). Recent work has shown that even in cases of high mobility (i.e., well-mixed communities), one can observe co-existence if higher-order interactions, such as antibiotic production and degradation, are considered (Kelsic et al., 2015). This is the case for *in vivo* experiments with bacterial colonies in the intestines of co-caged mice; these systems can be considered locally well-mixed and have high levels of coexistence (Kirkup and Riley, 2004).

High mutation rates, and processes such as horizontal gene transfer (HGT), have been long known to affect diversity in experimental microbial populations, especially during the initial phase of community assembly (Giraud et al., 2001; Dutta and Pan, 2002). But cyclic dominance models have tended to overlook the role of mutations and HGT in coexistence. Such studies have focused instead on detailed understanding of a small number of discrete species under a low mutation regime (Tarnita, 2018). Recently, studies have focused on understanding the effects of high mutational regimes in community assembly, which better represent microbial systems (Kotil and Vetsigian, 2018), and can generate frequent non-cyclic interactions (Park et al., 2020). Our efforts have considered an array of mutational regimes and characterized the importance of non-cyclic patterns of interactions in the assembly and maintenance of microbial communities.

Mutation size and *growth radius* (a proxy for mobility) were good predictors of SD mean and SD variance but yielded no discernable patterns in the phase space (**Figures 4.3D,E**). In contrast, community formation time (CFT) partitioned the space cleanly into four regions with distinct properties (**Figures 4.3F, 4.4A**). Communities that have low diversity (SD mean) and small fluctuations in diversity (i.e., low SD variance) (group 1) assemble quickly, and mobility

enhances diversity and co-existence in this scenario (**Figures 4.4A,B**). In contrast, mobility negatively affects diversity and coexistence for communities with high diversity (SD mean) and larger fluctuations (high SD variance) (group 2), and these communities assemble slower than group 1. For communities that take a long time to stabilize (groups 3 and 4), diversity is generally high (SD mean), but it fluctuates less and mobility has a negligible effect.

Mutation sizes were weakly correlated with CFT, as observed previously (Kotil and Vetsigian, 2018) but exhibited substantial variation in trends due to interactions with other model parameters. For low SD mean and variance (group 1) and high SD mean and variance (group 2), *mutation size* positively affects CFT (**Figure 4.4B**). This implies that communities with larger mutations take longer to stabilize, but only when the diversity outcome corresponds to these two groups (**Figure 4.4B**). If SD mean is high but SD variance is low, mutation affects CFT negatively, i.e., higher mutation helps communities converge faster (**Figure 4.4B**).

Community spatial structure is also important for coexistence (Szolnoki et al., 2014). However, rather than focusing on the final spatial structure, we examined the degree of spatial disturbance that occurred throughout the assembly process, which provides insight into the extent of mixing/migration that might occur in these communities. The categorization of spatial disturbance regimes was best predicted by species-level parameters (*kill margin* and *inhibit margin*), rather than by other spatial parameters in our model. This is somewhat surprising, but higher-order interactions are known to influence spatial patterns of coexistence (Kelsic et al., 2015). Importantly, we found a strong correspondence between the regimes of spatial disturbance and the groups produced by CFTs. The communities that formed quickly (group 1) tended to have low spatial disturbance (**Figure 4.4C**). Communities that took more time to stabilize and ended up

with high diversity through large overall diversity fluctuations (high SD mean and SD variance, group 2) have primarily a medium level of spatial disturbance. The communities that took longest to stabilize were characterized by high diversity (SD mean) and lower average fluctuations in diversity (SD variance) (groups 3 and 4) but have larger spatial disturbances.

All these observations point towards a complex interplay of mutation and mobility, which affects the assembly time of a community (CFT), and in turn controls the diversity of the assembled community. Although mobility (or more broadly, dispersal) is known to enhance diversity in ecological communities in both theoretical (Barton, et al., 2010; Tilman, 1994) and experimental (Howeth and Leibold, 2010; Venail et al., 2008) settings, especially at local scales (Cadotte, 2006), we found that mobility is linked to diversity only in communities with short and intermediate CFT (i.e., those that equilibrated relatively quickly). Mobility enhances diversity and coexistence in Group 1, where a subset of dominant (high relative abundance) species appear to set community dynamics. In this capacity, dispersal would appear to act in the classic disruptive fashion, permitting coexistence where it would otherwise not occur (Hassell et al., 1991; Snyder et al., 2003). In contrast in Group 2, where communities are characterized by both high diversity and high fluctuations, mobility has a negative effect on diversity and coexistence, in keeping with the capacity for dispersal to homogenize otherwise diverse systems (McKinney and Lockwood, 1999; Pauchard and Shea, 2006). This dichotomy, together with the absence of an important role for mobility in Group 4, where communities are characterized by long-term nonequilibrium dynamics, offers an intriguing target for future integrative research.

The mechanistic light shed upon these patterns by studying the Shannon equitability index (which measures the distribution of relative abundance of species in a community) and Simpson's index

(which measures the distribution of ‘dominance’ of species in a community) is also worth mentioning. The more equitable a community (higher Shannon equitability), the lower the chance of a particular species dominating the interactions (lower value of Simpson’s index) and therefore, the longer it takes for the community dynamics to stabilize (longer CFTs), as seen in the behavior of the four groups (see **Figures 4.4D-G**). The importance of species dominance to the maintenance of diversity in this model, which is structured via non-resource based antagonistic interactions, is intriguingly similar to the critical role that species dominance plays in both real and theoretical communities structured via resource-based competition (Holt et al., 1994; Sasaki et al., 2011; Seabloom et al., 2003; Stanley Harpole et al., 2006).

Collectively, these results demonstrate the rich spatiotemporal dynamics that are possible when large numbers of microbial species with limited but heterogeneous rules for aggressive, inhibitory, and vulnerable interactions live in a common space. Although these findings are focused primarily on non-resource-based antagonistic interactions between microbes, such dynamics may also be relevant for coral communities and other spatially structured systems featuring diverse types of interspecific interactions (see Laird and Schamp, 2018; Karlson and Jackson, 1981; Precoda et al., 2017).

Our results also point towards the important yet often-overlooked role that transient dynamics play in the behavior and structure of ecological systems (see Hastings et al., 2018). Usually, models of ecological dynamics use asymptotically stable behavior or values at stability to explore patterns in the system (Hastings et al., 2018). Admittedly, we have followed this approach in our exploration of how different parameters affect microbial communities’ long-run behavior (i.e., after CFT is reached). In addition, however, we also focused on categorizing spatio-temporal disturbance

regimes and connecting these transient phenomena to system diversity. Great opportunities exist for future work investigating temporal diversity dynamics and how they relate to spatial heterogeneity and system processes. Such investigations can shed light on how transient microbial dynamics and assembly history affects the species diversity seen in microbial communities (Fukami, 2004).

Even though our model introduced the use of continuous axes for species traits and interactions, we note that the *post-hoc* binning procedure creates a bridge back to the matrix/network framework that has characterized community ecology studies for decades (Allesina and Tang, 2015; Butler and O'Dwyer, 2018, 2020; May, 1972; McCann, 2000; Sinha and Sinha, 2005). In this case, however, a multi-layered network perspective would be necessary to accommodate the different kinds of antibiotic-mediated interactions (production, inhibition, and vulnerability). Future work could also incorporate other important forms of microbial interactions (e.g., mutualism, cross-feeding, resource competition) into the continuous species framework we have developed. Such investigations would allow exploration of how the interplay between resource-based and antagonistic interactions jointly shape diversity dynamics. However, to do this we need to understand the interrelationships between resource-based and antagonistic interactions across species (see Gonzalez and Mavridou, 2019; Peterson et al., 2020; West et al., 2006). Such investigations could be highly beneficial by providing a more complete view of how higher-order, intransitive interactions shape community assembly, stability, and diversity in natural systems.

Methods

Model Description

Here we outline an algorithmic presentation of the model. The parameters of the model and their corresponding symbols are detailed in **Table 4.1**. The simulation takes place on a 200 x 200 square lattice with toroidal boundary conditions. This size of this simulation was a tradeoff between being small enough for computational feasibility and large enough to allow interesting dynamics. We found 200 x 200 to be a reasonable range for our computation budget. The model of microbial dynamics on this simulation is split into two parts: the microbial (phenotypic) space and the environment space (see **Figure 4.1**). This division of the model is purely for convenience and can easily be reinterpreted where the microbes and the environment are not differentiated. For the microbial phenotypic space, each cell in the lattice represents a single ‘variant’ of microbe and is parameterized by a stateful three-dimensional species vector $\mathbf{R}^3 \in \{\emptyset, [0, 1]^3\}$, where $\emptyset$ represents a cell that is unoccupied and $[0, 1]$ is a bounded real value that defines the parameters of any occupants. The three elements of the species vector are the species’ vulnerability (S_v), the species’ inhibition value (S_i), and the species’ antibiotic production value (S_a). In the microbial space, neighboring cells do not interact with each other directly, but instead affect the environment space which then affects the microbial space. The environment space is stateless and instead always resets to the most recent effects from the microbial space, or in other words, no antibiotics remain in the environment between timesteps. This can just be interpreted as diffusion on a less granular timescale. A cell in the environment space is parameterized by a single scalar value $\mathbf{R}^1 \in \{\emptyset, [0, 1]\}$, which represents the antibiotic occupying that cell space (C_a), or indicates the cell contains no antibiotics, represented by $\emptyset$.

Parameter	Type	Definition
<i>Species antibiotic production value</i> (S_a)	Species-level, Species-specific	Defines what species are vulnerable to the antibiotic that an individual produces
<i>Species vulnerability value</i> (S_v)	Species-level, Species-specific	Defines what antibiotics (and hence species) an individual is vulnerable to
<i>Species inhibition value</i> (S_i)	Species-level, Species-specific	Defines the species whose antibiotics an individual can inhibit
<i>Kill radius</i> (K_r)	Spatial, Global	Maximum spatial distance of diffusion of the antibiotics produced by an individual
<i>Inhibit radius</i> (I_r)	Spatial, Global	Maximum spatial distance of diffusion of the inhibitors to antibiotics of produced by an individual
<i>Growth radius</i> (G_r)	Spatial, Global	Maximum spatial distance at which an individual can reproduce by producing a copy of itself with certain possible mutations
<i>Kill margin</i> (ϵ_k)	Species-level, Global	The total span centered around the specified <i>Species antibiotic production value</i> (S_a), which can be affected/killed by the antibiotic of a given species
<i>Inhibit margin</i> (ϵ_i)	Species-level, Global	The total span centered around the specified <i>Species inhibition value</i> (S_i), whose antibiotic can be inhibited by a given species
<i>Mutation size</i> (μ)	Species-level, Global	Maximum value by which the species-specific parameters can be additively altered during reproduction/copying

Table 4.1: Brief description of model parameters

Note that our model involves only a phenotypic space for microbial traits. There is no underlying genotypic space, and our phenotypic microbial space abstracts away genotypic changes into expressed phenotypic traits. Doing this in a meaningful way is a very difficult problem, and this is especially true for bacterial toxin-based antagonisms, where the basic nature of the interactions remains at a preliminary level of research and discovery (García-Bayona et al., 2018; Gorter et al., 2021; Granato et al., 2019; Heilbronner et al., 2021; Pérez-Gutiérrez et al., 2013; Peterson et al., 2020; Schoustra et al., 2012; Sibinelli-Sousa et al., 2021; Ting et al., 2018; Wang et al., 2018).

At fixed-interval time steps, the state of the microbial space is updated synchronously in four stages. First, an antibiotic release step occurs, which simulates the process of microbes diffusing antibiotics into the environment. Here each cell's antibiotic value (C_a), in the environment space is randomly assigned a species' antibiotic value (S_a) of one other cell from the microbial space that is within the antibiotic radius (K_r). Note that rewrites of the environment cell are possible during this step, so the order of cell updates is randomized. Second, an inhibition step occurs which handles any inhibiting effects of the antibiotics in the environment space. In this step, for every cell in the microbe space, an inhibition radius (I_r) around it is checked for other cells in the environment space which contain an antibiotic value (C_a) such that the inhibition value is within a marginal range of the antibiotic value, $I_v - \epsilon_i \leq C_a \leq I_v + \epsilon_i$. If this is the case, the cell in the environment space antibiotic value (C_a) is made empty, $\emptyset$. Third, a kill step occurs where we iterate over every cell, comparing the antibiotic value (C_a) in the environment space and the individual vulnerability value (S_v) in the microbial space. If the individual vulnerability value (S_v) is within the range of the antibiotic value (C_a), $C_a - \epsilon_a \leq S_v \leq C_a + \epsilon_a$ then the individual is killed and the cell in the microbial space is made empty, $\emptyset$. Fourth and last, a growth step occurs,

where for every cell in the microbial space that it is not empty, another cell in the microbial space that is within that cell's growth radius (G_r) is randomly selected. If the selected cell is empty, the species vector of the growing microbe cell is copied to the selected cell. The copied parameters are also further modified according to a mutation size (μ). This modification is a non-correlated additive perturbation of the species' vector by a uniform value sampled from $U(-\mu, \mu)$ and the value is restricted in the range $[0,1]$ by adding or subtracting 1 if it is outside this range. This ensures that all species have equal likelihood of interactions with other species by putting the species parameters on a three-dimensional toroid. From a theoretical perspective, this growth can be seen as a model of growth probability, where the probability that an individual will reproduce is equal to $E(C)/F(C)$ where $E(C)$ is the number of empty cells within radius G_r of the current C and $F(C) = \text{floor}(2\pi G_r^2)$ is equal to the total number of cells within radius G_r of cell C .

Model Assumptions

There are three major assumptions/rules underlying our model implementation. The first is that the model only attends to toxin-based antagonistic interactions. Little evidence is available to characterize the relationship between antagonistic interactions and other forms of microbial interactions (e.g., mutualisms, resource competition) across species (Heilbronner et al., 2021; Peterson et al., 2020). Therefore, our aim here was to build a 'base' model that allows for the investigation of antagonistic interactions with more nuance and across more dimensions of complexity than was possible in previous work on the subject. The interplay between antagonistic interactions and other types of microbial interactions may be explored in future work as more information becomes available.

The second assumption pertains to the structure of the microbial trait space. Our model represents species traits of ‘antibiotic production’, ‘vulnerability’, and ‘inhibition’ as continuous values on a real number line. A chief advantage of this approach is that a continuous trait space makes it easy to introduce mutations that preserve some history of the evolutionary process. Though critical to representations of species evolution, such mutations are difficult to combine with graph-based models of species interactions because it is unclear how mutation would update the graph edges (Kelsic et al., 2015). We define the presence or absence of a species interaction using a simple thresholded Euclidean distance (termed a ‘margin’ in our model). In a multi-species context, the larger the margin, the greater the connectivity of the graph of interacting species, and likewise the smaller the margin, the sparser the graph.

These distance-based margins assume that the trait space is a uniform, three-dimensional toroid, which has implications for how we deal with changes in trait space. In particular, changing the value of one trait in one species and a second trait in another species by the same value d , will result in the interacting edge between those two species remaining unchanged in the graph of interacting species. For example, if species A is vulnerable to species B’s antibiotic, and the trait values of vulnerability (S_v) in A and antibiotic production (S_a) in B each mutate by a value d , species A will remain vulnerable to species B’s antibiotic. This approach, which is mathematically simple compared to alternative representations of trait evolution (see Supplemental Methods and **Fig. S4.6**), is a reasonable starting point because the true nature of the relationship between antibiotic production and vulnerability across microbial species is very much unknown (see García-Bayona et al., 2018; Gorter et al., 2021; Granato et al., 2019; Heilbronner et al., 2021; Pérez-Gutiérrez et al., 2013; Peterson et al., 2020; Schoustra et al., 2012; Sibinelli-Sousa et al., 2021; Ting et al., 2018; Wang et al., 2018).

In real species, a small genotypic change may or may not correspond to a small change in phenotypic space, and some phenotypes can be independently reached via alternative sets of genotypic mutational processes (including horizontal gene transfers). To accommodate such phenomena, models of microbial diversity and / or mutational processes often adopt an explicitly phenotypic modeling approach (see Kelsic et al., 2015, Vetsigian et al., 2011). The phenotypic structure of our model allows for such complexity—including the possibility of both small and large phenotypic changes in a single generation, and no specific correlation exists between distance on the phenotypic space along any axis and the (implicit, unmodeled) genetic change from which it results. Modeled in this fashion, phenotypic changes could plausibly represent the outcome of a diverse set of microbial mutational processes, given that the details of such processes affecting antagonistic microbial interactions are still in their infancy (for details see García-Bayona et al., 2018; Gonzalez and Mavridou, 2019; Gorter et al., 2021; Granato et al., 2019; Heilbronner et al., 2021; Pérez-Gutiérrez et al., 2013; Peterson et al., 2020; Schoustra et al., 2012; Sibinelli-Sousa et al., 2021; Ting et al., 2018; Wang et al., 2018). We emphasize that we do not model mutations at a genotype level because horizontal gene transfer, which plays a major role in trait evolution in microbes, especially bacteriocins (see García-Bayona et al., 2018; Heilbronner et al., 2021; Peterson et al., 2020), is not easily amenable to such modeling.

For our third assumption, we stipulate that, within a run of our model, mutations only affect species identity (i.e., parameters S_a, S_v, S_i within the continuous trait space). Other parameters (i.e., those governing spatial aspects of interactions, the species-breadth (“margin”) of interactions, and mutation size; see **Table 4.1**) remain constant in a given run. We make this assumption because antagonistic spatial interactions between bacteria (which we emulate via parameters for kill radius and inhibit radius) are typically affected by diffusion properties of the toxin compounds, which

have similar physical properties within a given class of molecules but can vary across classes of toxins (Peterson et al., 2020). Hence, for computational purposes, we keep the radius parameters constant over a run. Likewise, we keep the margin parameters constant over a run because they govern the generality of the interactions within the fixed kill and inhibit radii. We keep the maximum mutation size constant within runs to facilitate comparisons across different regimes of trait alteration.

Model Implementation

We implemented the simulation in *C++* using the *SFML* graphics library for rendering. It is possible to parallelize the updating of sub-steps of the cells due to the synchronous nature of the model, but we chose to use a simpler sequential algorithm that looped through each cell one at a time. We did, however, parallelize the running of multiple simulations when sweeping parameter spaces. This allowed us to reduce the total computation time required by two orders of magnitude. Each simulation was run for 2000 timesteps or until stability/stagnation (also termed the Community Formation Time (CFT)). Stagnation/stability was determined by comparing the state of all cells between two consecutive time steps, and if it remained unchanged for a total of five-time steps, the simulation was terminated. This helped to quickly remove a large number of simulations that terminated early due to poor parameter settings. Because the simulation was so large, collecting per-cell data across many parameters was not practical. Alternatively, we chose to compute population-level statistics online during each simulation using Shannon Diversity mean and variance (over time) and then from the analysis of this lower-granularity data, we selected individual parameter settings to explore population-level statistics. Further, we also built a qualitative application to allow quick classification of simulations, assigning them to a class based

on their spatial heterogeneity/disturbance dynamics. Simulations were assigned to one of three classes according to how the spatial dynamics evolved (**Fig. 4.2**).

To track population-level statistics, it was necessary to discretize the continuous species space. l -discretization (or l -binning) assigned individuals to the same species if the three parameters of those individuals had the same integer values after performing a binning transformation on them like so: $\left\lfloor \frac{S_v}{l} \right\rfloor, \left\lfloor \frac{S_i}{l} \right\rfloor, \left\lfloor \frac{S_a}{l} \right\rfloor$. When recording the population-level data, we stored the number of individuals per species every time step, using a standard 0.05-discretization. Although the choice of 0.05-discretization is arbitrary, we emphasize that the binning only affects the resolution of the statistical analysis and has no influence on the dynamics of the model itself. We found 0.05 to be a small enough value for which the simulation statistics did not blend away, but not too small such that there were no measurable statistics (see Data Analysis below). For comparison, we repeated the complete set of analyses on system properties at three other reasonable species bin sizes (0.03, 0.07, 0.09) (see **Fig. S4.5** for more details).

For every parameter setting, we repeated the simulation ten times, seeding the pseudo-random number generator with a unique value for each repetition. Every simulation began by assigning every cell vector the all-empty state $(-1, -1, -1, -1, -1)$ except the center cell which is assigned a random value in the range $[-1, 1]$ for the first three elements of its vector. This initial cell represented a randomly created individual. One potential issue was that if the center cell was assigned a random value such that it was vulnerable to its own antibiotic, it would kill itself and the simulation would terminate immediately. Although this was possible, it was rarely the case, and it did not significantly affect our results due to our use of multiple runs per parameter setting.

Data Analysis

We ran a total of 10.49 million runs of our model with various parameter initializations, and performed data analysis with *Python 3.7* and the *R* statistical language. For these runs, we obtained a manual-classification of the size of the spatial disturbance (low, medium and high). The simulations which stabilized quickly (fewer than 20 time-steps) with no further changes in spatial patterns and the ones that had (more-or less) stabilized by 20 time-steps, except for a small number of point fluctuations in the grid, were categorized as the ‘low disturbance’ regime (**Figure 4.2 A1**). The simulations which, had not stabilized after 20 time-steps but exhibited small waves of diversity replacements which diminished over time were termed as ‘medium’ disturbance. All medium disturbance simulations stabilized before 500 time-steps (**Figure 4.2 B1**). The remaining simulations did not stabilize even after 500 time-steps, nor had reduced amplitude of waves of diversity replacements; these were categorized as the ‘high’ spatial disturbance regime (**Figure 4.2 C1**).

We also calculated the mean Shannon diversity, Shannon equitability, and Simpson’s index over time (until a community stabilized, i.e., reached the community formation time, CFT), and the variance in each of the three metrics over the same time period. The disturbance classification, SD mean, and SD variance were our three outcomes of interest and are the focus of the results. We later connect the results to the values of Shannon equitability and Simpson’s index.

To understand how parameters affect diversity dynamics, we determined the partial rank correlation coefficient (PRCC) of various parameters with respect to the three outcomes of interest, using the *sensitivity* package (Iooss et al., 2021) in R. We also performed Random Forest (RF) regressions and classification, using the *randomForest* package (Liam and Wiener, 2002) in R, to

identify which parameters were the strongest predictors of the patterns in predicting the outcomes of interest. We optimized the number of parameters available for splitting at each tree node in the RF using out-of-bag error (OOB) (Liam and Wiener, 2002). Please note that the overlay figures (**Figures 4.3 D-F, S4.2**) have been plotted with pruned samples for better visualization.

For clustering values on the SD mean versus SD variance space with respect to CFT, we used k-means clustering in *R* and optimized for the number of clusters using the elbow method, the D index, and the Hubert indices (which all yielded similar results) using the package *NbClust* (Charrad et al., 2014). For each obtained cluster in the Shannon diversity space, we performed a random forest analysis to calculate differences in variable importance scores across these clusters. For each cluster, we also performed PRCC and random forest analyses, in order to delineate the effects that individual parameters had on the different subsets of the diversity space.

Acknowledgements

NSF DMS-1853465 (to WFF) and the University of Maryland supported this work. We would like to thank Brantley Hall, Stephanie A. Yarwood, and Jake L. Weissman for their insights and discussions that made the work better. We would also like to thank the Santa Fe Institute (SFI) Complex Systems Summer School (CSSS) for making this collaboration possible. A.S. would like to acknowledge NSF award DGE-1632976 for training and intellectual support.

Code and Data Availability

All code and required data to replicate our results are available in the following repository:

<https://github.com/levifussell/MicroEvo>.

References

1. Abrudan, M.I., Smakman, F., Grimbergen, A.J., Westhoff, S., Miller, E.L., Van Wezel, G.P. and Rozen, D.E., 2015. Socially mediated induction and suppression of antibiosis during bacterial coexistence. *Proceedings of the National Academy of Sciences*, 112(35), pp.11054-11059.
2. Allesina, S. and Tang, S., 2015. The stability–complexity relationship at age 40: a random matrix perspective. *Population Ecology*, 57(1), pp.63-75.
3. Bairey, E., Kelsic, E.D. and Kishony, R., 2016. High-order species interactions shape ecosystem diversity. *Nature communications*, 7(1), pp.1-7.
4. Barton, A.D., Dutkiewicz, S., Flierl, G., Bragg, J. and Follows, M.J., 2010. Patterns of diversity in marine phytoplankton. *Science*, 327(5972), pp.1509-1511.
5. Butler, S. and O’Dwyer, J.P., 2018. Stability criteria for complex microbial communities. *Nature communications*, 9(1), pp.1-10.
6. Butler, S. and O’Dwyer, J.P., 2020. Cooperation and stability for complex systems in resource-limited environments. *Theoretical Ecology*, 13(2), pp.239-250.
7. Cadotte, M.W., 2006. Dispersal and species diversity: a meta-analysis. *The American Naturalist*, 167(6), pp.913-924.
8. Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A., 2014. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61, pp.1-36.
9. Chase, J.M., 2003. Community assembly: when should history matter?. *Oecologia*, 136(4), pp.489-498.

10. Chesson, P., 2000. Mechanisms of maintenance of species diversity. *Annual review of Ecology and Systematics*, 31(1), pp.343-366.
11. Cordero, O.X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., Le Roux, F., Mincer, T. and Polz, M.F., 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science*, 337(6099), pp.1228-1231.
12. Coyte, K.Z., Schluter, J. and Foster, K.R., 2015. The ecology of the microbiome: networks, competition, and stability. *Science*, 350(6261), pp.663-666.
13. Czárán, T.L., Hoekstra, R.F. and Pagie, L., 2002. Chemical warfare between microbes promotes biodiversity. *Proceedings of the National Academy of Sciences*, 99(2), pp.786-790.
14. Durrett, R. and Levin, S., 1998. Spatial aspects of interspecific competition. *Theoretical population biology*, 53(1), pp.30-43.
15. Dutta, C. and Pan, A., 2002. Horizontal gene transfer and bacterial diversity. *Journal of biosciences*, 27(1), pp.27-33.
16. Frey, E., 2010. Evolutionary game theory: Theoretical concepts and applications to microbial communities. *Physica A: Statistical Mechanics and its Applications*, 389(20), pp.4265-4298.
17. Friedman, J., Higgins, L.M. and Gore, J., 2017. Community structure follows simple assembly rules in microbial microcosms. *Nature ecology & evolution*, 1(5), pp.1-7.
18. Fukami, T., 2004. Assembly history interacts with ecosystem size to influence species diversity. *Ecology*, 85(12), pp.3234-3242.
19. García-Bayona, L. and Comstock, L.E., 2018. Bacterial antagonism in host-associated microbial communities. *Science*, 361(6408).

20. Ghoul, M. and Mitri, S., 2016. The ecology and evolution of microbial competition. *Trends in microbiology*, 24(10), pp.833-845.
21. Ginzburg, L.R., Akçakaya, H.R. and Kim, J., 1988. Evolution of community structure: competition. *Journal of theoretical biology*, 133(4), pp.513-523.
22. Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M. and Taddei, F., 2001. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *science*, 291(5513), pp.2606-2608.
23. Gonzalez, D. and Mavridou, D.A., 2019. Making the best of aggression: the many dimensions of bacterial toxin regulation. *Trends in microbiology*, 27(11), pp.897-905.
24. Gorter, F.A., Manhart, M. and Ackermann, M., 2020. Understanding the evolution of interspecies interactions in microbial communities. *Philosophical Transactions of the Royal Society B*, 375(1798), p.20190256.
25. Gorter, F.A., Tabares-Mafla, C., Kassen, R. and Schoustra, S.E., 2021. Experimental Evolution of Interference Competition. *Frontiers in microbiology*, 12, p.602.
26. Granato, E.T., Meiller-Legrand, T.A. and Foster, K.R., 2019. The evolution and ecology of bacterial warfare. *Current biology*, 29(11), pp.R521-R537.
27. Green, J. and Bohannan, B.J., 2006. Spatial scaling of microbial biodiversity. *Trends in ecology & evolution*, 21(9), pp.501-507.
28. Grilli, J., Barabás, G., Michalska-Smith, M.J. and Allesina, S., 2017. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666), pp.210-213.
29. Gross, K., Cardinale, B.J., Fox, J.W., Gonzalez, A., Loreau, M., Wayne Polley, H., Reich, P.B. and van Ruijven, J., 2014. Species richness and the temporal stability of biomass

- production: a new analysis of recent biodiversity experiments. *The American Naturalist*, 183(1), pp.1-12.
30. Guill, C. and Drossel, B., 2008. Emergence of complexity in evolving niche-model food webs. *Journal of theoretical biology*, 251(1), pp.108-120.
 31. Hall, J.P., Brockhurst, M.A. and Harrison, E., 2017. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1735), p.20160424.
 32. Hassell, M.P., Comins, H.N. and Mayt, R.M., 1991. Spatial structure and chaos in insect population dynamics. *Nature*, 353(6341), pp.255-258.
 33. Hastings, A., 2001. Transient dynamics and persistence of ecological systems. *Ecology Letters*, 4(3), pp.215-220.
 34. Hastings, A., Abbott, K.C., Cuddington, K., Francis, T., Gellner, G., Lai, Y.C., Morozov, A., Petrovskii, S., Scranton, K. and Zeeman, M.L., 2018. Transient phenomena in ecology. *Science*, 361(6406).
 35. Heilbronner, S., Krismer, B., Brötz-Oesterhelt, H. and Peschel, A., 2021. The microbiome-shaping roles of bacteriocins. *Nature Reviews Microbiology*, pp.1-14.
 36. Hibbing, M.E., Fuqua, C., Parsek, M.R. and Peterson, S.B., 2010. Bacterial competition: surviving and thriving in the microbial jungle. *Nature reviews microbiology*, 8(1), pp.15-25.
 37. Higgins, L.M., Friedman, J., Shen, H. and Gore, J., 2017. Co-occurring soil bacteria exhibit a robust competitive hierarchy and lack of non-transitive interactions. *bioRxiv*, p.175737.

38. Holt, R.D., Grover, J. and Tilman, D., 1994. Simple rules for interspecific dominance in systems with exploitative and apparent competition. *The American Naturalist*, 144(5), pp.741-771.
39. Howeth, J.G. and Leibold, M.A., 2010. Species dispersal rates alter diversity and ecosystem stability in pond metacommunities. *Ecology*, 91(9), pp.2727-2741.
40. Hutchinson, G.E., 1961. The paradox of the plankton. *The American Naturalist*, 95(882), pp.137-145.
41. Iooss, B., et al., 2020. sensitivity: Global Sensitivity Analysis of Model Outputs. R package version 1.20.0. <https://CRAN.R-project.org/package=sensitivity>
42. Jeraldo, P., Sipos, M., Chia, N., Brulc, J.M., Dhillon, A.S., Konkel, M.E., Larson, C.L., Nelson, K.E., Qu, A., Schook, L.B. and Yang, F., 2012. Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences*, 109(25), pp.9692-9698.
43. K. Vetsigian, R. Jajoo, R. Kishony, Structure and evolution of *Streptomyces* interaction networks in soil and in silico. *PLoS Biol*, 9(10), e1001184 (2011).
44. Karlson, R.H. and Jackson, J.B., 1981. Competitive networks and community structure: a simulation study. *Ecology*, 62(3), pp.670-678.
45. Kelsic, E.D., Zhao, J., Vetsigian, K. and Kishony, R., 2015. Counteraction of antibiotic production and degradation stabilizes microbial communities. *Nature*, 521(7553), pp.516-519.
46. Kerr, B., Neuhauser, C., Bohannan, B.J. and Dean, A.M., 2006. Local migration promotes competitive restraint in a host–pathogen 'tragedy of the commons'. *Nature*, 442(7098), pp.75-78.

47. Kerr, B., Riley, M.A., Feldman, M.W. and Bohannan, B.J., 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature*, 418(6894), pp.171-174.
48. Kirkup, B.C. and Riley, M.A., 2004. Antibiotic-mediated antagonism leads to a bacterial game of rock–paper–scissors in vivo. *Nature*, 428(6981), pp.412-414.
49. Kotil, S.E. and Vetsigian, K., 2018. Emergence of evolutionarily stable communities through eco-evolutionary tunnelling. *Nature ecology & evolution*, 2(10), pp.1644-1653.
50. Laird, R.A. and Schamp, B.S., 2018. Calculating competitive intransitivity: Computational challenges. *The American Naturalist*, 191(4), pp.547-552.
51. Le Roux, F. and Blokesch, M., 2018. Eco-evolutionary dynamics linked to horizontal gene transfer in *Vibrios*. *Annual review of microbiology*, 72, pp.89-110.
52. Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. *R News* 2(3), 18—22.
53. Libby, E., Hébert-Dufresne, L., Hosseini, S.R. and Wagner, A., 2019. Syntrophy emerges spontaneously in complex metabolic systems. *PLoS computational biology*, 15(7), p.e1007169.
54. Loeuille, N. and Loreau, M., 2005. Evolutionary emergence of size-structured food webs. *Proceedings of the National Academy of Sciences*, 102(16), pp.5761-5766.
55. Machado, D., Maistrenko, O.M., Andrejev, S., Kim, Y., Bork, P., Patil, K.R. and Patil, K.R., 2021. Polarization of microbial communities between competitive and cooperative metabolism. *Nature ecology & evolution*, 5(2), pp.195-203.
56. May, R.M., 1972. Will a large complex system be stable?. *Nature*, 238(5364), pp.413-414.
57. McCann, K.S., 2000. The diversity–stability debate. *Nature*, 405(6783), pp.228-233.

58. McKinney, M.L. and Lockwood, J.L., 1999. Biotic homogenization: a few winners replacing many losers in the next mass extinction. *Trends in ecology & evolution*, 14(11), pp.450-453.
59. Momeni, B., Xie, L. and Shou, W., 2017. Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *Elife*, 6, p.e25051.
60. O. Diekmann, 2001. A beginners guide to adaptive dynamics. Summer school on mathematical biology, pp.63-100.
61. Park, H.J., Pichugin, Y. and Traulsen, A., 2020. Why is cyclic dominance so rare?. *Elife*, 9, p.e57857.
62. Pauchard, A. and Shea, K., 2006. Integrating the study of non-native plant invasions across spatial scales. *Biological invasions*, 8(3), pp.399-413.
63. Pérez-Gutiérrez, R.A., López-Ramírez, V., Islas, A., Alcaraz, L.D., Hernández-González, I., Olivera, B.C.L., Santillán, M., Eguiarte, L.E., Souza, V., Travisano, M. and Olmedo-Alvarez, G., 2013. Antagonism influences assembly of a *Bacillus* guild in a local community and is depicted as a food-chain network. *The ISME journal*, 7(3), pp.487-497.
64. Perlin, M.H., Clark, D.R., McKenzie, C., Patel, H., Jackson, N., Kormanik, C., Powell, C., Bajorek, A., Myers, D.A., Dugatkin, L.A. and Atlas, R.M., 2009. Protection of *Salmonella* by ampicillin-resistant *Escherichia coli* in the presence of otherwise lethal drug concentrations. *Proceedings of the Royal Society B: Biological Sciences*, 276(1674), pp.3759-3768.
65. Peterson, S.B., Bertolli, S.K. and Mougous, J.D., 2020. The central role of interbacterial antagonism in bacterial life. *Current Biology*, 30(19), pp.R1203-R1214.

66. Precoda, K., Allen, A.P., Grant, L. and Madin, J.S., 2017. Using traits to assess nontransitivity of interactions among coral species. *The American Naturalist*, 190(3), pp.420-429.
67. Raynaud, X. and Nunan, N., 2014. Spatial ecology of bacteria at the microscale in soil. *PloS one*, 9(1), p.e87217.
68. Reichenbach, T., Mobilia, M. and Frey, E., 2007. Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. *Nature*, 448(7157), pp.1046-1049.
69. Rodríguez-Verdugo, A. and Ackermann, M., 2021. Rapid evolution destabilizes species interactions in a fluctuating environment. *The ISME Journal*, 15(2), pp.450-460.
70. Sasaki, T. and Lauenroth, W.K., 2011. Dominant species, rather than diversity, regulates temporal stability of plant communities. *Oecologia*, 166(3), pp.761-768.
71. Schoustra, S.E., Dench, J., Dali, R., Aaron, S.D. and Kassen, R., 2012. Antagonistic interactions peak at intermediate genetic distance in clinical and laboratory strains of *Pseudomonas aeruginosa*. *BMC microbiology*, 12(1), pp.1-9.
72. Seabloom, E.W., Harpole, W.S., Reichman, O.J. and Tilman, D., 2003. Invasion, competitive dominance, and resource use by exotic and native California grassland species. *Proceedings of the National Academy of Sciences*, 100(23), pp.13384-13389.
73. Shores, N., Hegreness, M. and Kishony, R., 2008. Evolution exacerbates the paradox of the plankton. *Proceedings of the National Academy of Sciences*, 105(34), pp.12365-12369.
74. Sibilini-Sousa, S., Hespanhol, J.T. and Bayer-Santos, E., 2021. Targeting the Achilles' heel of bacteria: different mechanisms to break down the peptidoglycan cell wall during bacterial warfare. *Journal of Bacteriology*, 203(7), pp.e00478-20.

75. Sinha, S. and Sinha, S., 2005. Evidence of universality for the May-Wigner stability theorem for random networks with local dynamics. *Physical Review E*, 71(2), p.020902.
76. Snyder, R.E. and Chesson, P., 2003. Local dispersal can facilitate coexistence in the presence of permanent spatial heterogeneity. *Ecology letters*, 6(4), pp.301-309.
77. Stanley Harpole, W. and Tilman, D., 2006. Non-neutral patterns of species abundance in grassland communities. *Ecology letters*, 9(1), pp.15-23.
78. Szolnoki, A., Mobilia, M., Jiang, L.L., Szczesny, B., Rucklidge, A.M. and Perc, M., 2014. Cyclic dominance in evolutionary games: a review. *Journal of the Royal Society Interface*, 11(100), p.20140735.
79. Tarnita, C.E., 2018. Fast evolution unlocks forbidden communities. *Nature ecology & evolution*, 2(10), pp.1525-1526.
80. Tilman, D., 1994. Competition and biodiversity in spatially structured habitats. *Ecology*, 75(1), pp.2-16.
81. Ting, S.Y., Bosch, D.E., Mangiameli, S.M., Radey, M.C., Huang, S., Park, Y.J., Kelly, K.A., Filip, S.K., Goo, Y.A., Eng, J.K. and Allaire, M., 2018. Bifunctional immunity proteins protect bacteria against FtsZ-targeting ADP-ribosylating toxins. *Cell*, 175(5), pp.1380-1392.
82. Van Hoek, M.J. and Merks, R.M., 2017. Emergence of microbial diversity due to cross-feeding interactions in a spatial model of gut microbial metabolism. *BMC systems biology*, 11(1), pp.1-18.
83. Venail, P.A., MacLean, R.C., Bouvier, T., Brockhurst, M.A., Hochberg, M.E. and Mouquet, N., 2008. Diversity and productivity peak at intermediate dispersal rate in evolving metacommunities. *Nature*, 452(7184), pp.210-214.

84. Wang, C.Y., Patel, N., Wholey, W.Y. and Dawid, S., 2018. ABC transporter content diversity in *Streptococcus pneumoniae* impacts competence regulation and bacteriocin production. *Proceedings of the National Academy of Sciences*, 115(25), pp.E5776-E5785.
85. Weber, M.F., Poxleitner, G., Hebisch, E., Frey, E. and Opitz, M., 2014. Chemical warfare and survival strategies in bacterial range expansions. *Journal of The Royal Society Interface*, 11(96), p.20140172.
86. West, S.A., Griffin, A.S., Gardner, A. and Diggle, S.P., 2006. Social evolution theory for microorganisms. *Nature reviews microbiology*, 4(8), pp.597-607.
87. Wright, E.S. and Vetsigian, K.H., 2016. Inhibitory interactions promote frequent bistability among competing bacteria. *Nature communications*, 7(1), pp.1-7.

Supplementary materials

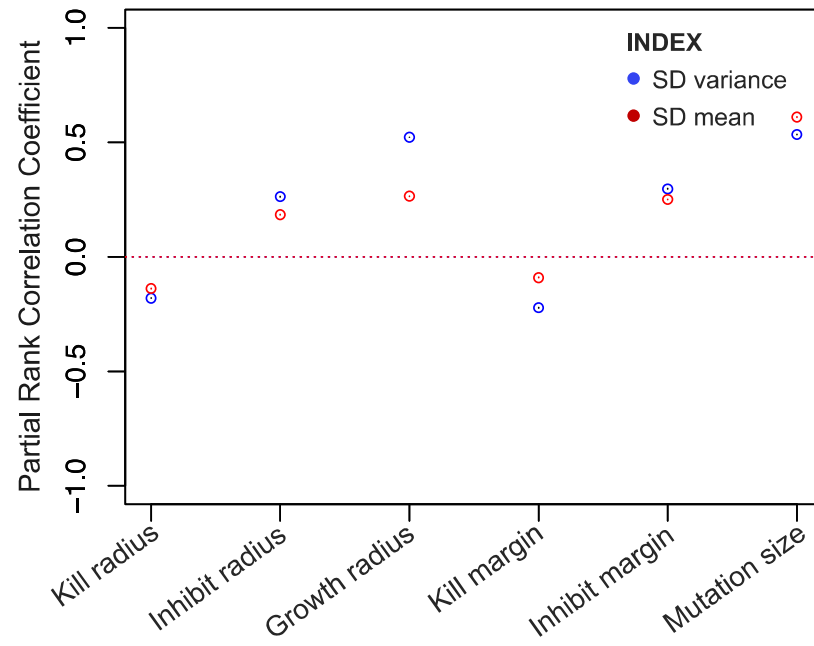


Figure S4.1: Partial rank correlation coefficients (PRCC) of model parameters in relation to the Shannon Diversity (SD) mean and variance. Please note that different parameters have similar values of PRCC for both the diversity metrics, and the greatest difference in the trends of the two metrics was for *growth radius*.

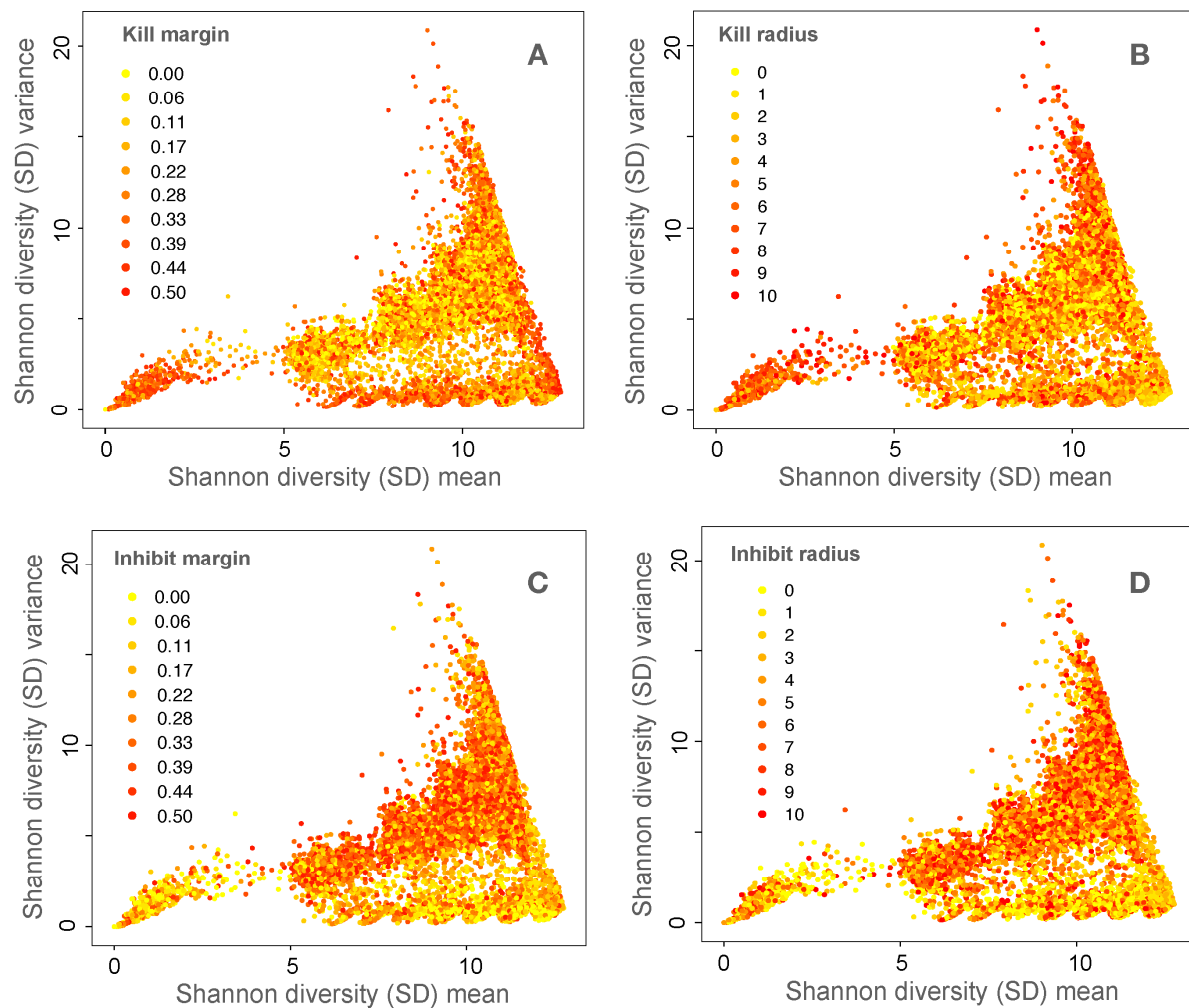


Figure S4.2: Overlaying values of parameters, (A) *Kill margin*, (B) *Kill radius*, (C) *Inhibit margin*, and (D) *Inhibit radius*, over the respective values on Shannon Diversity (SD) mean and SD variance space. No obvious patterns are discernable in how these parameters affect the diversity space as compared to the ones in Figure 3 D-F, where distinct patterns in clustering of the diversity space can be observed.

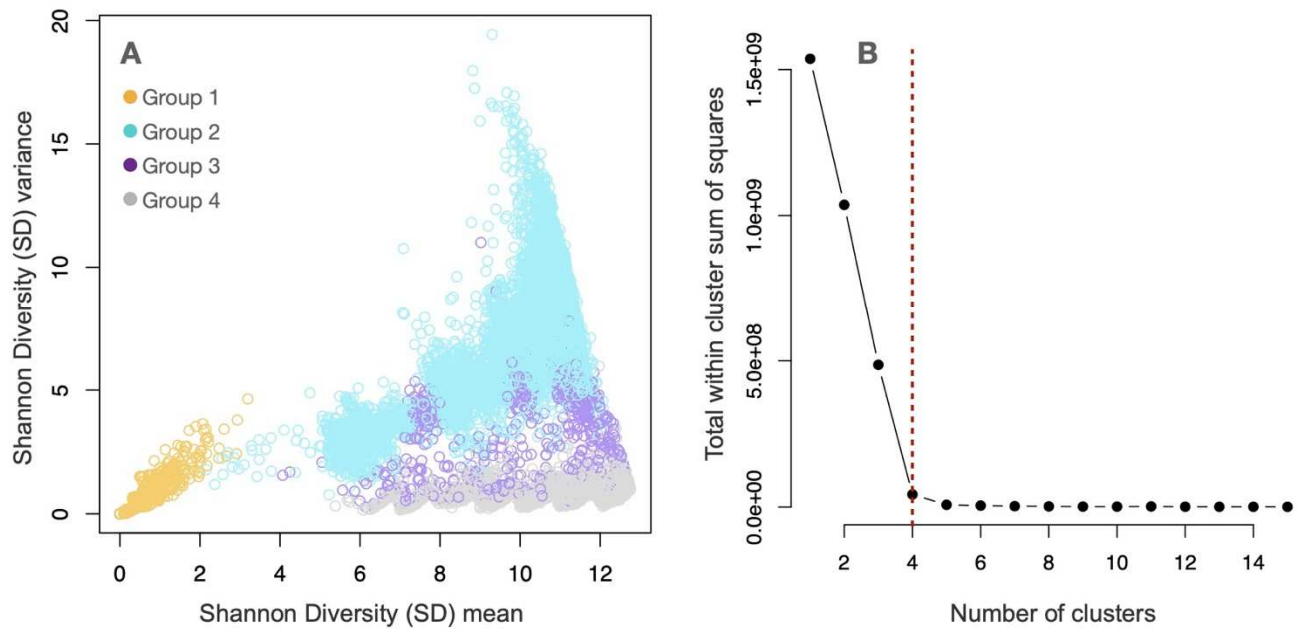


Figure S4.3: (A) k-means clustering of the Shannon Diversity (SD) mean-variance space by the community formation time (CFT) value. (B) shows the optimal number of clusters using the elbow method. We obtained similar results when optimizing clusters using the D index and the Hubert indices.

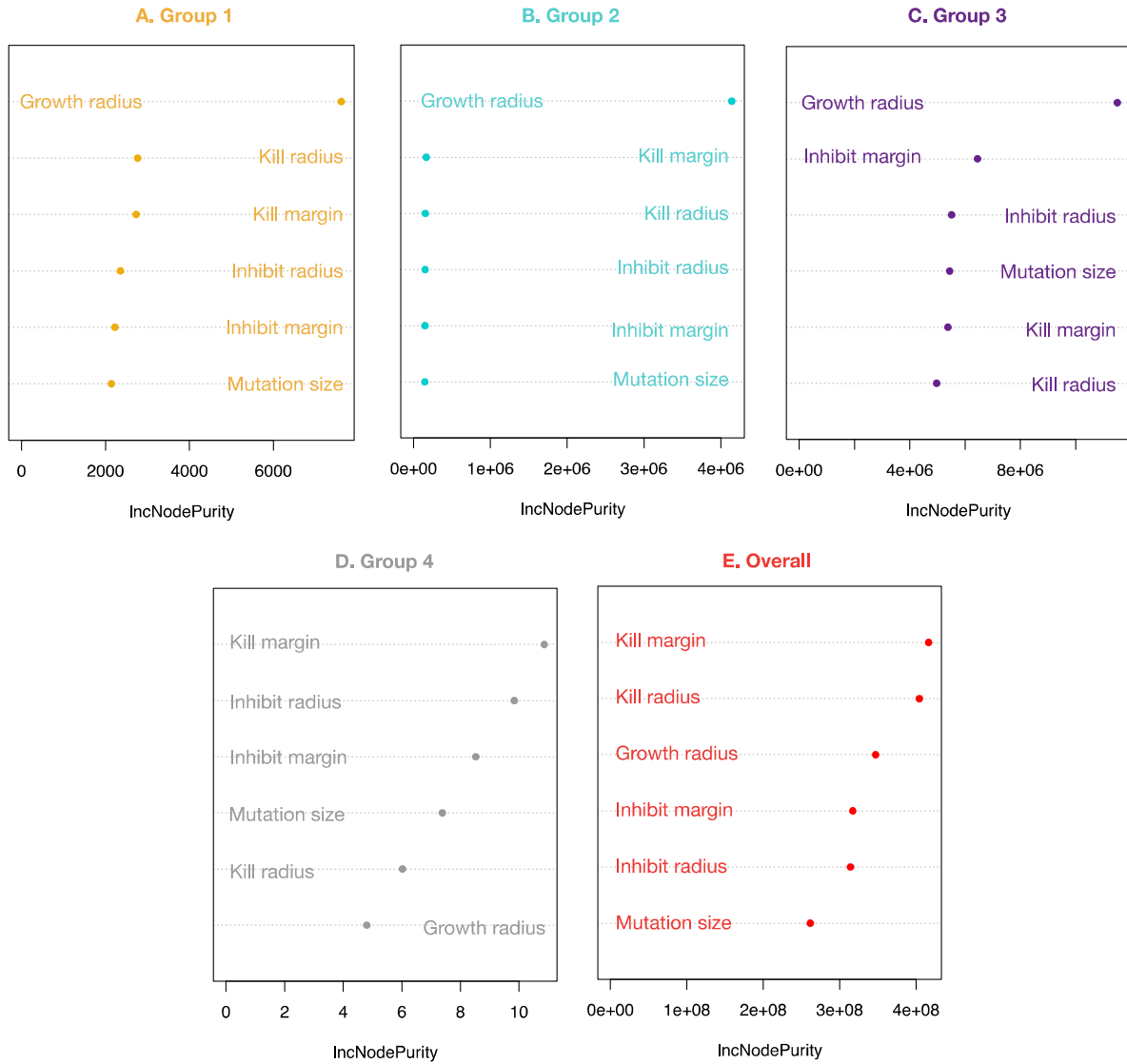


Figure S4.4: Random forest regressions (10,000 trees each with optimization for splits using out-of-bag (OOB) error) for community formation time (CFT) for the four diversity groups clustered using k-means. Variance explained for- Group 1: 30.2%, Group 2: 90.6%, Group 3: 27.3%, Group 4: 6.0%, Overall: 22.3%. Note: IncNodePurity refers to the total decrease in node impurities from splitting on a given parameter, averaged over all trees.

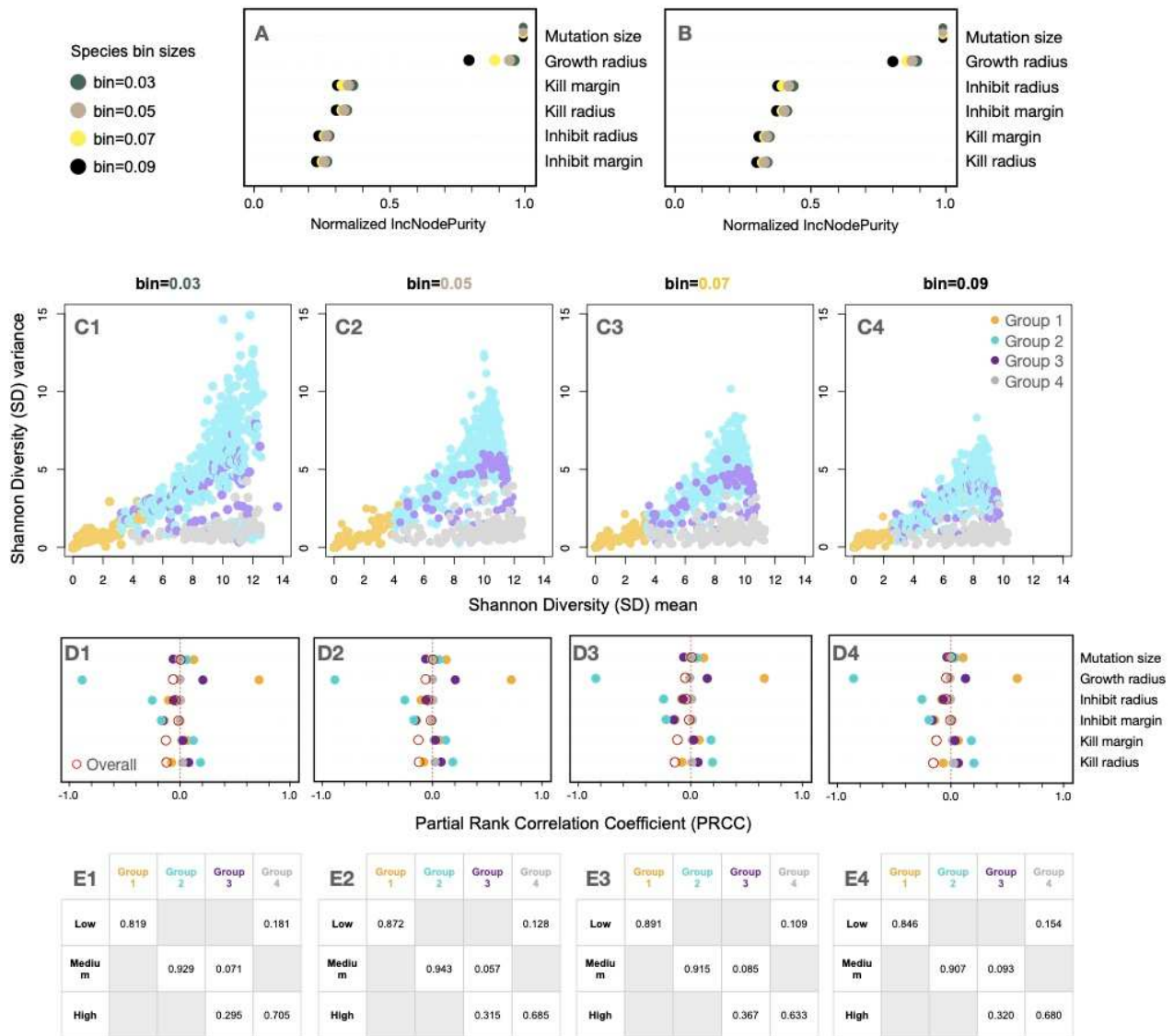


Figure S4.5: Dependence of the results on species bin sizes. (A) and (B) depict the normalized variable importance scores for a random forest regression of Shannon diversity (SD) mean and variance, respectively, at the four bin sizes. Percent variance explained remained between 68-75% for all bin sizes, and all other results were similarly consistent. (The only thing to note is that the dependence on growth radius (a mobility-related metric) weakens with increasing bin size, which may occur because increasing bin size results in greater discretization for a given length scale). (C1-4) depict the k-means clustering of the space of SD mean and variance for the four bin sizes. Increasing bin size collapses the range of SD mean and SD variance of the resultant communities, but otherwise no changes are obvious. (D1-4) depict the Partial Rank Correlation Coefficients (PRCC) of variables on Community Formation Time (CFT), which are consistent across bin sizes. (E1-4) depict how the spatial categorizations correspond to the diversity groups at four bin sizes, and again the results are consistent.

Supplemental Methods. Species interactions can be generalized by keeping the distance-based threshold function but changing the geometric representation of the trait space by defining some function $f(x): \mathbf{R}^3 \rightarrow \mathbf{R}^3$ which maps the uniform trait space to a *transformed trait space*. The threshold function (margin) is then applied to this transformed trait space to determine the interactions between species. In other words, if we imagine the interactions in our model to be represented by a graph (as in the case of Kelsic et al. 2015), the original trait space defines the identity of the species (i.e., nodes in the graph) and the transformed trait space (after being acted upon by the threshold function) defines the interactions (i.e., the edges in the graph).

Our current model simply uses the identity mapping $f(x) = x$, and therefore the transformed trait space and the original trait space are equivalent. The identity map is the simplest functional assumption. Instead of the uniform toroidal space in our current model, other trait space geometries are possible which could represent very intricate non-linear relationships between how individual species, and their traits, interact and evolve (See Figure S6 for simplified examples). The nature of these interactions is not explicitly known from current experimental literature and there is much debate around it, i.e., there is a dearth of empirical information on the structure and dependencies of these traits (e.g., whether they are independent, or constrained; see Schoustra et al., 2012; Pérez-Gutiérrez et al., 2013; Gonzalez et al., 2019; Granato et al., 2019; Peterson et al., 2020; Gorter et al., 2021). In the absence of any guiding information, we assumed an independent, uniform linear space of microbial traits for simplicity. This leaves room to develop the model to more complex definitions of trait interactions based on future experimental findings.

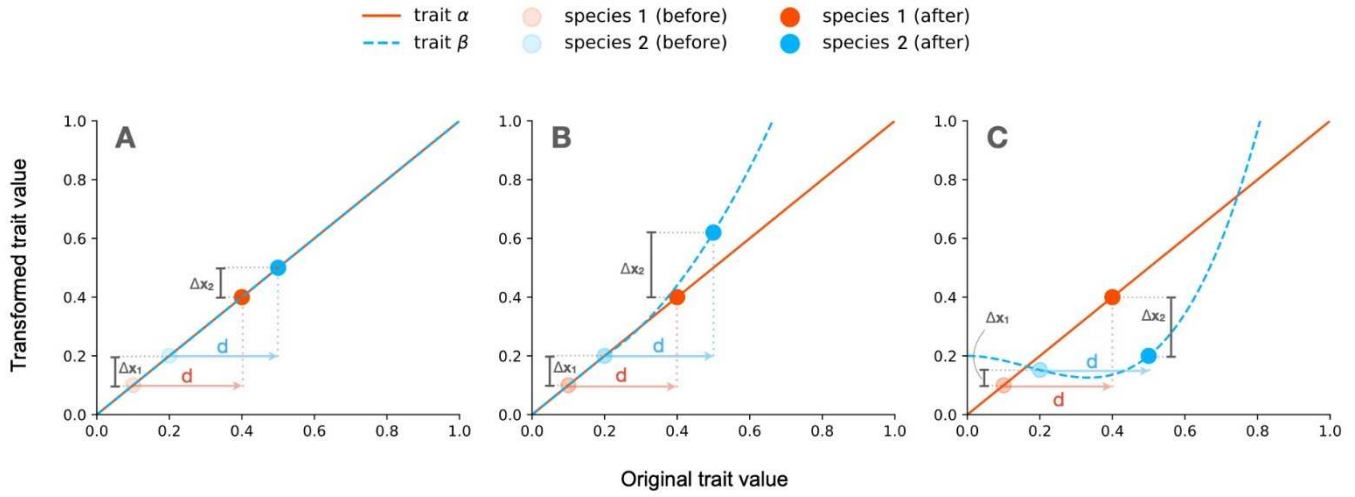


Figure S4.6: Simple example of two-trait space: (A) depicts the distance between the original trait α for species 1 and trait β for species 2 before and after both traits are mutated by a value ' d '. The transforming function (i.e., the function that transforms the trait value for use in interaction mapping) is linear (i.e., identity function, as is the case with our model). The trait distance between both species is the same before and after mutation ($\Delta x_1 = \Delta x_2$). These principles of linearity and equidistance underpin the simplicity of our approach to trait evolution. (B) depicts the same species and traits, but trait β has a different geometry (here the transforming function is exponential) such that mutating both traits by a value ' d ' does not preserve the trait distance between the two species ($\Delta x_1 \neq \Delta x_2$). (C) is similar to (B) but with a more non-linear transforming function. Both (B) and (C) represent more complex (and more specific) assumptions about trait interrelationships than the scenario in (A) which we adopted. If we imagine the interactions in our model to be represented by a graph (as in the case of Kelsic et al. 2015), the original trait space (X-axis) defines the identity of the species (i.e., nodes in the graph) and the transformed trait space (Y-axis) (after being acted upon by the threshold function) defines the interactions (i.e., the edges in the graph).

Chapter 5 : Exploring noise, degeneracy, and determinism in biological networks with the ‘einet’ package

Please note that the contents of this chapter have been published as:

Klein*, B., Swain*, A., Byrum, T., Scarpino, S.V. and Fagan, W.F., 2022. Exploring noise, degeneracy, and determinism in biological networks with the `einet` package. *Methods in Ecology and Evolution*.
<https://doi.org/10.1111/2041-210X.13805>

(*equal contributions; first co-authors)

Abstract

1. Understanding noise in networks and finding the right scale to represent a system are important problems in network biology. Most research focuses on the raw, micro-scale network from data/simulations and seldom explores the scale-dependence of properties.
2. Here, we introduce the `einet` package, which looks at the most informative scale in a biological network using recent concepts from information theory and network science.
3. `einet` uses two metrics: Effective information, which measures the interplay between degeneracy and determinism in a network’s edges, and causal emergence, which finds the scale of the network with the highest effective information.
4. `einet` is available in R and Python and provides tools to explore noise and scale dependency in networks as well as compare information flow and noise across networks.

Introduction

The functioning of biological systems depends heavily on the interaction of its constituents at various levels of organization, be it at the subcellular level or at the ecosystem scale. In order to

understand the complexity of such interactions in a more abstract yet workable way, people have turned to networks as a useful representation (Gosak et al., 2018; Gysi et al., 2020). The use of network science and its tools in biology has exploded in the past two decades, owing to the burst of high-resolution data and improved computational capabilities. The application of network science has improved our fundamental understanding of many biological systems and structures, such as food webs (Dunne et al., 2002; Pascual & Dunne, 2006; Shaw et al., 2021), neuronal functioning and signaling (Sporns, 2014; Bassett & Sporns, 2017), gene interactions and regulation (Constanzo et al., 2019), mutualistic interactions (Vázquez et al., 2009), epidemiological networks (Zang et al., 2015), microbial interactions (Faust & Raes, 2012), and animal social networks (Pinter-Wollman et al., 2014).

Most network studies focus on the structure and/or the dynamics of the constructed network itself, and probe how its various properties can explain or predict certain biological activity. However, sometimes due to noise—either inherent to the system or introduced via measurement and observation—the structure and dynamics of networks can be affected by the way the networks are constructed (Tsimring, 2014). The kind of noise introduced through observation will often be specific to particular systems under study. Past work has focused on the estimation, and in rare cases, the correction of such noise (Newman, 2018; Freilich et al., 2020; Swain et al., 2021). Beyond such estimation, it can often be difficult to deal with the noise that is inherently present in complex biological systems regardless of whether the noise is from measurement or is inherent to the system in question.

Because of the high levels of noise and stochasticity associated with biological systems, and because such systems are often subjected to random fluctuations and perturbations, biological systems tend to be redundant and degenerate in their functioning. This has led to a lot of research

on the resilience and robustness of biological systems through degeneracy and other similar features (Tononi et al., 1999; Edelman & Gally, 2001). These studies have changed our view about the function of degeneracy in such systems, emphasizing increased resilience to sudden perturbations, and functional equivalence of degenerate entities (see Edelman & Gally, 2001, Ahn et al., 2006, Hoel et al., 2020). Research on biological networks often seeks to connect certain deterministic or partially deterministic processes through the lens of redundancy, resilience, and degeneracy (Ahn et al., 2006). Indeed, the interplay between complexity and degeneracy/determinism in biological systems has received a great deal of attention in network science and the complex systems literature (Hoel et al., 2013).

Generally, network methods focus on the literal constructed network to explore the effects of noise, perturbations, degeneracy and deterministic processes—but such an approach might not be able to tease apart certain informative aspects of the structure of the network (Klein & Hoel, 2020). Most biological networks have a certain form of hierarchy, and it is quite possible that despite the noise at the micro-scale (or the literal network), a network can possess more informative structure at a different scale (Klein & Hoel, 2020; Hoel et al., 2020). The big question, then, is how to infer these higher, content-rich scales and to decide which scale is the most informative to use in analyses.

The concepts of *effective information* and *causal emergence* are network metrics devised to confront these problems (Klein & Hoel, 2020; Hoel et al., 2020). These two indices, which are based on information theory (Hoel et al., 2013), measure the information associated with the structure of a network, which can be used to identify the presence of informative higher scales. In this work, we briefly describe the metrics, their interpretations and usage, through implementation both as an R package and a Python library.

Overview of the metrics

The deterministic nature of a network can be measured in terms of the uncertainty in the interactions among the nodes. That is, if activity on a given node influences or induces changes in another node's behavior, we can study the inherent uncertainty in this relationship. This influence may come in the form of neuronal activity, where one neuron's firing could induce firing in a neighboring neuron; in social networks, influence could come in the form of a connection strength, where my activity may depend more on a close friend than a mere acquaintance; or in biological systems, where the co-expression of two biological processes may indicate a dependency between the two processes.

If node j is connected to node i , and the behavior of i suddenly changes, what are the chances that node j 's behavior is also changed? To quantify this, we can calculate the average Shannon entropy of the out- and in-edges of nodes in a network. The entropy (or uncertainty) of out-edges corresponds to the degree of *determinism* in the network connections, while *degeneracy* captures how evenly the in-edges are distributed across the network (see following section). Effective information (EI) brings these two concepts (i.e., *determinism*, *degeneracy*) together, and is defined as the difference between the entropy of the average out-weight in the network minus the average entropy of the out-weights (see Klein & Hoel, 2020, for a detailed discussion):

$$EI = H(\langle W_i^{out} \rangle) - \langle H(W_i^{out}) \rangle \quad \dots (5.1)$$

where H refers to the entropy, and $\langle \rangle$ is the average over all nodes, denoted by i . Effective information varies with network size, so researchers typically use a size-normalized measure known as *effectiveness* to characterize networks:

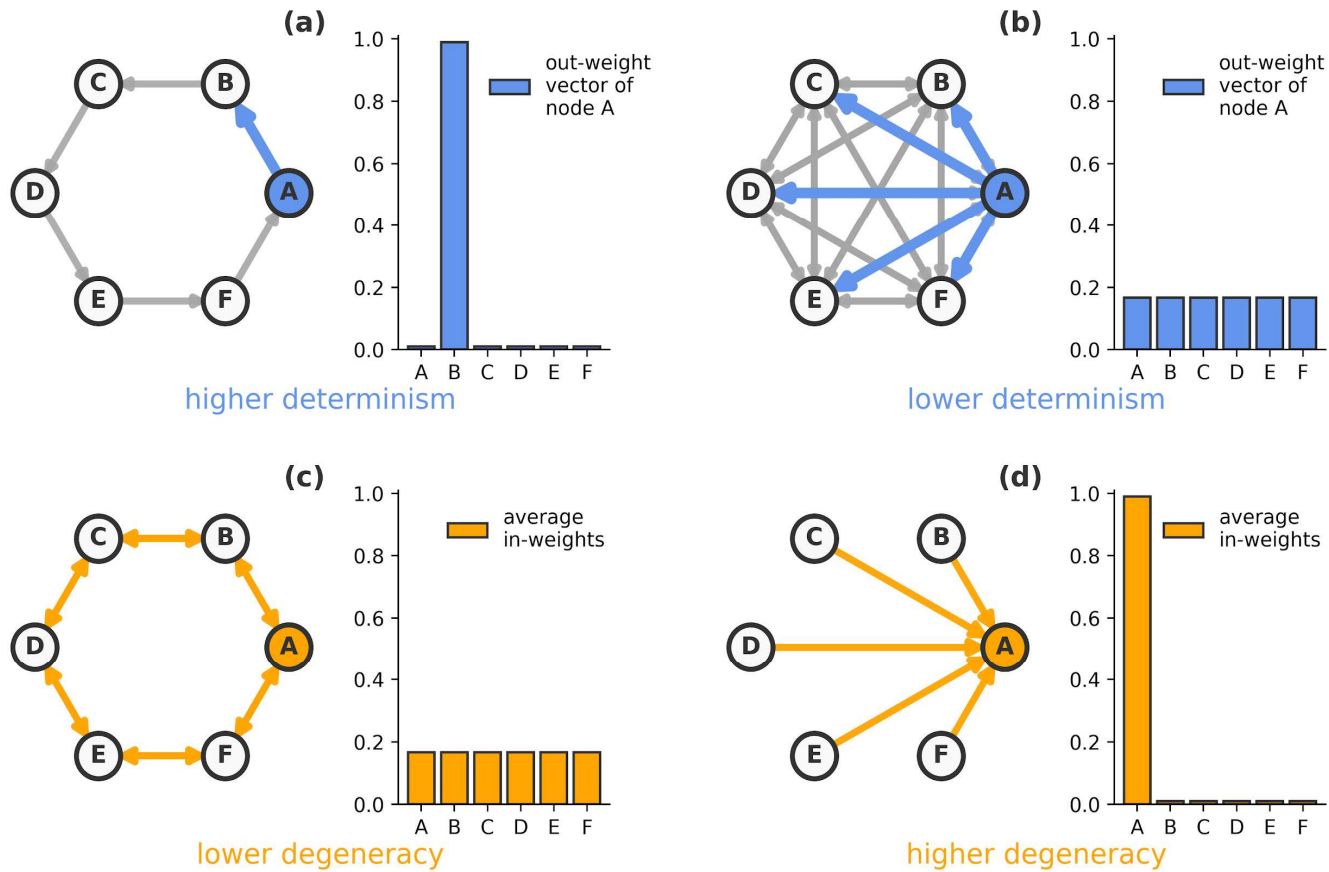


Figure 5.1: Determinism and degeneracy. Simple networks highlighting the differences between high/low determinism and degeneracy. **(a)** In this network, random walkers on node A will traverse to node B with probability of 1.0. **(b)** In this network, random walkers on node A will traverse to any other node in the network with probability $1/n$, where n is the size of the network. **(c)** In this network, no single node has a higher in-weight than any other node, which means that degeneracy has been minimized. **(d)** In this network, every node has a single outgoing edge, each of which connects to node A; this is an example of a maximally degenerate network

$$Effectiveness = \frac{EI}{\log_2(n)} \quad \dots (5.2)$$

where EI is normalized with respect to the number of nodes, n . Using these measures, we can compare different networks, highlighting structural properties that are associated with more or less uncertainty in the edges of a network. This gives researchers a lens through which to view network dynamics, robustness, and growth, and can inform why we may expect different systems across nature and society to vary in the amount of noise contained in their connectivity structures. For example, researchers have studied the effectiveness of various biological or socio-technical systems (Klein & Hoel, 2020), protein interaction networks (Hoel et al., 2020), animal social networks (Swain, Williams, Di Felice & Hobson, 2021; van der Marel et al., 2021), and more.

Breaking down effective information: Determinism and degeneracy

As previously mentioned, the EI is defined by two key components: *determinism* and *degeneracy*. Determinism refers to the average uncertainty over outgoing edges in a network. If a given node i has a single outgoing edge, there is no uncertainty about where a random walker on node i will traverse to next—in other words, node i increases the network’s *determinism*. On the other hand, if i has outgoing edges to the other n nodes in a network, we are maximally uncertain about the subsequent location of the random walker (**Figures 5.1a, 5.1b**). Whereas determinism is a measure that quantifies uncertainty about the *future* states of random walkers in the network, *degeneracy* captures uncertainty about the source of the random walkers on a given node. Consider a node with only one incoming edge compared to a node with n incoming edges; when few nodes have disproportionately many connections, the network becomes more uncertain, on average, about past

states of random walkers (see **Figures 5.1c, 5.1d**). Note: the notion of random walkers is a useful—if incomplete—way of formalizing flow or transitions in networks. Because networks are such a ubiquitous tool for representing complex systems, there are a number of ways to represent dynamical processes on networks from different domains; random walks have the benefit of being well-studied mathematically and are general enough such that they add few domain-specific assumptions. The prospect of re-defining a version of the effective information metric by replacing random walk dynamics with domain-specific dynamics is a rich and promising future direction of research.

Different networks display different behavior when it comes to *determinism* and *degeneracy*. For example, networks with heavy-tailed degree distributions, which are common across nature and society, typically have higher *degeneracy* (and lower *EI* as a result). Similarly, networks containing dense subgraphs and/or community structure typically have lower *determinism* (and lower *EI* as a result). Networks that are sparse and have low variance in degree have higher *EI* and have less uncertainty about the micro-scale interactions that take place in the network.

In practice, determinism and degeneracy may correspond to different facets of biological systems. For example, a highly deterministic relationship in a gene expression network would correspond to almost one gene almost certainly causing the expression of another. A degenerate structure in, for example, an ecological network would be more *sensitive*, a connectivity structure where fluctuations in the prevalence of any of a number of species would influence the prevalence of a target species.

Causal emergence: Informative higher scales

Networks that have higher determinism and lower degeneracy are less noisy and have more certainty in the interactions among their nodes; in this sense they are more effective (i.e., they have higher effectiveness). One recent development in network science and information theory has been the introduction of causal emergence to the study of complex networks (Klein & Hoel, 2020); this is a phenomenon where the original micro-scale network is recast as a coarse-grained, macro-scale network so as to increase the effective information of the original network. Algorithmically, this involves selecting a subgraph of nodes and their edges—a subgraph with low determinism and/or high degeneracy—from the original network and replacing the subgraph with a single macro node. A core functionality of the `einet` package is that it automates this coarse-graining process, using a variety of techniques from spectral graph theory to find and isolate regions of low determinism and/or high degeneracy. Users simply input a network into the `causal_emergence` function, and if there are highly noisy subgraphs, the function will use those subgraphs to create a higher scale macro representation of the network. We illustrate this process in Figure 2, using the protein-protein interaction network of *Mycoplasma putrefaciens* as an example.

Causal emergence (the quantity) is defined as the difference between the EI of the original, micro-scale network and the EI of a coarse-grained, higher scale representation of the network. In the context of ecological or biological systems, causal emergence of a network can be interpreted in many ways. For example, in insect social networks, we might expect to see more informative higher scales that capture more of the emergent coordination of the collective as a whole (e.g. in ant colonies; see Swain et al., 2021).

Higher order networks are of growing interest to network scientists, and they appear to be relatively ubiquitous, appearing in a variety of contexts from biology to economics to social

structures and more. What causal emergence adds to this growing body of work is the ability to quantify why certain higher scale structures emerge, identifying common structural properties at the micro scale that tend to facilitate higher order structure. How to interpret a given higher order structure (i.e., output from the `einet` package) can vary greatly across different disciplines, but put simply: causal emergence uses the pattern of noise/uncertainty at the micro scale to define the most informative scale at which to represent a complex system. These higher scale structures may hold theoretical import (e.g. by comparing the amount of causal emergence across different species groups or ecosystems, as in Klein, Hoel, et al., 2020), or they help uncover measurement noise that must be addressed before studying networks from a given system (similar to recent advances in network permutation testing, as in Puga-Gonzalez, Sueur, & Sosa, 2021). Future work may also explore the use of causal emergence in network comparison (Hartle et al., 2020) or parameter selection when modeling ecological networks (e.g., animal social networks, species interaction networks, etc.), in order to constrain the structure of networks produced by a generative model. Ultimately a comprehensive, prescriptive description of how to interpret higher order structure in networks requires more theoretical development; we view `einet` as important to such future theoretical advances.

Implementation and usage

These tools are available through `einet`, both as an R package (GitHub) and a Python library (Github). For details about the implementation of the spectral algorithm for computing the micro-to-macro scale mappings, see (Griebenow et al., 2019).

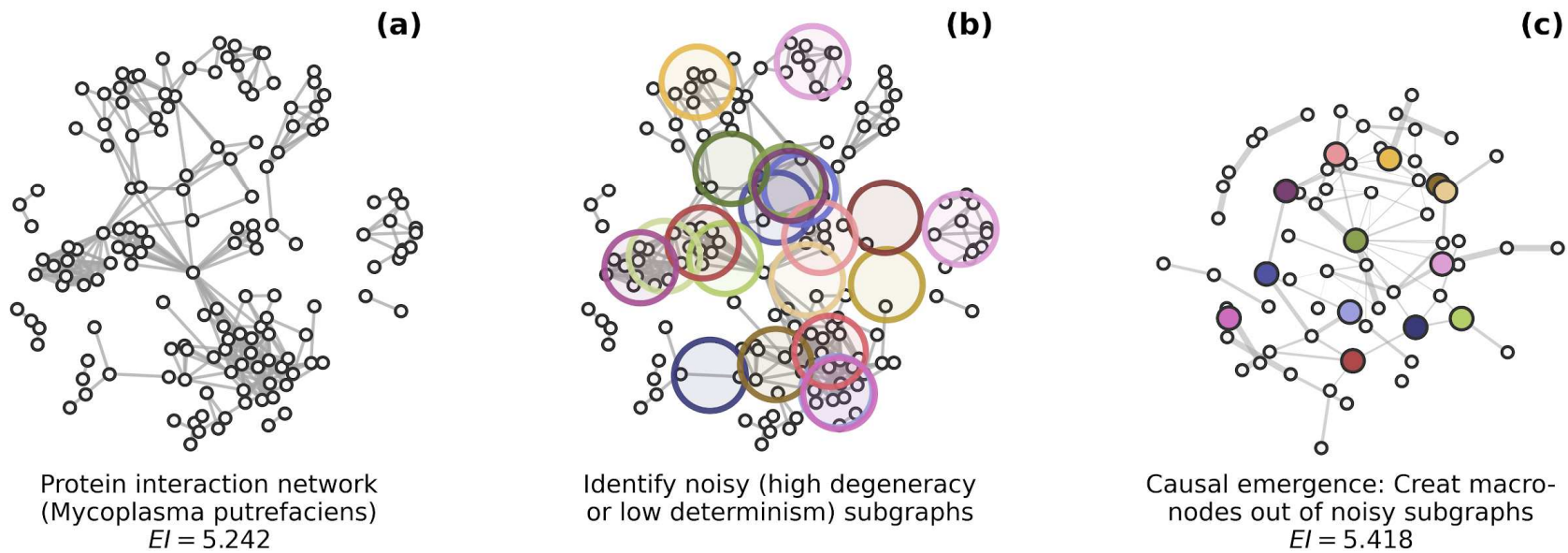


Figure 5.2: Causal emergence in protein networks. To illustrate the basic steps for calculating causal emergence in networks, we use the protein-protein interaction network of *Mycoplasma putrefaciens* as an example. **(a)** The original (micro scale) network. **(b)** The different color circles are meant to highlight subgraphs in the original network that are especially noisy—either because they consist of densely connected nodes or because they are sources of degeneracy. **(c)** In order to calculate causal emergence, we construct a macro scale network by grouping the subgraphs from (b) into macro nodes, while preserving the total in/out edge weights into the selected subgraph. Note here that the EI of the network in (a) and the EI from (c) are different: the macro scale network has more effective information than the original micro scale network.

Example usage (R)

Effective information and effectiveness

```
>>> library(devtools)
>>> library(igraph)
>>> install_github(`travisbyrum/einet`)
>>> library(einet)
>>> G <- karate
>>> ei_micro <- effective_information(G)
>>> sprintf("EI micro: %.6f",ei_micro)
EI micro: 2.350095
>>> eff_micro <- ei_micro / log2(gorder(G))
>>> sprintf("Effectiveness: %.6f",eff_micro)
Effectiveness: 0.461939
```

Causal emergence

```
>>> library(einet)
>>> library(igraph)
>>> set.seed(2)
>>> G <- karate
>>> CE <- causal_emergence(G)
>>> ei_macro <- CE$ei_macro
>>> sprintf("EI macro: %.6f",ei_macro)
EI macro: 2.410400
```

The CE object returned by the `causal_emergence` function is a list object composed of `g_micro` (original network), `g_macro` (coarse-grained network), `mapping` (a list mapping from the micro nodes to their corresponding macro node, after coarse-graining happened), `ei_macro` (effective information of the coarse-grained network), `ei_micro` (original network's effective information), `effectiveness` (original network's effective information, divided by $\log(n)$ (normalization), where n is the number of nodes) and `ce` (the normalized difference in effective information between the micro and macro scale networks). Because the `causal_emergence()` function depends on the initial conditions of the search for the higher scale mapping, we recommend running the function several times and selecting the mapping with the largest effective information (note that this is the source of the small differences between the causal emergence in the R implementation and the Python below).

Example usage (Python)

Effective information and effectiveness

```
>>> from ei_net import effective_information
>>> import networkx as nx; import numpy as np
>>> G = nx.karate_club_graph()
>>> EI_micro = effective_information(G)
>>> print('EI micro: %.6f'%EI_micro)
EI micro: 2.350095
```

Causal emergence

```
>>> from ei_net import causal_emergence
>>> import networkx as nx; import numpy as np
>>> np.random.seed(2)
```

```

>>> CE = causal_emergence(G)
>>> EI_macro = CE['EI_macro']
>>> print('EI macro: %.6f'%EI_macro)
EI macro: 2.415379
>>> N = G.number_of_nodes()
>>> eff_gain = (EI_macro-EI_micro)/np.log2(N)
>>> print('Effectiveness gain: %.6f"%eff_gain)
Effectiveness gain: 0.012832

```

The CE object returned by the `causal_emergence` function is a dictionary with the following keys: `G_micro` (original network), `G_macro` (coarse-grained network), `mapping` (a list mapping from the micro nodes to their corresponding macro node, after coarse-graining happened), `EI_macro` (effective information of the coarse-grained network), and `EI_micro` (original network's effective information). In the Python implementation of this software, there is an additional `causal_emergence_spectral()` function that leverages tools from spectral graph theory to compute causal emergence, and does so much faster than the greedy search used in `causal_emergence()` (see Griebenow, Klein, & Hoel, 2019 for more details).

Software and Code Availability

The R package is available at github.com/travisbyrum/einet. The Python code is available at github.com/jkbren/einet; this repository also includes introductory Jupyter notebooks detailing various aspects of effective information, causal emergence, and networks.

Acknowledgements

BK and SS acknowledge the support of a grant from the John Templeton Foundation (61780). The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. AS thanks National Science Foundation award DGE-1632976 for training and support.

References

1. Ahn, A.C., Tewari, M., Poon, C.S., & Phillips, R.S., 2006. The limits of reductionism in medicine: could systems biology offer an alternative?. *PLoS Med*, 3(6), e208.
2. Bassett, D.S. & Sporns, O., 2017. *Network Neuroscience*. *Nature Neuroscience*, 20(3), 353-364.
3. Byrum, T., Swain, A., Klein, B., & Fagan, W.F., 2020. einet: Effective Information and Causal Emergence. R package version 0.1.0. CRAN.R-project.org/package=einet
4. Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., & Andrews, B., 2019. Global genetic networks and the genotype-to-phenotype relationship. *Cell*, 177(1), 85-100.
5. Dunne, J.A., Williams, R.J., & Martinez, N.D., 2002. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20), 12917-12922.
6. Edelman, G.M., & Gally, J.A., 2001. Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, 98(24), 13763-13768.
7. Faust, K. & Raes, J., 2012. Microbial interactions: From networks to models. *Nature Reviews Microbiology*, 10(8), 538-550.
8. Freilich, M.A., Rebolledo, R., Corcoran, D., & Marquet, P.A., 2020. Reconstructing ecological networks with noisy dynamics. *Proceedings of the Royal Society A*, 476(2237), 20190739.
9. Gosak, M., Markovič, R., Dolenshek, J., Rupnik, M.S., Marhl, M., Stožer, A., & Perc, M., 2018. Network science of biological systems at different scales: A review. *Physics of Life Reviews*, 24, 118-135.

10. Griebenow, R., Klein, B., & Hoel, E., 2019. Finding the right scale of a network: Efficient identification of causal emergence through spectral clustering. *arXiv preprint arXiv:1908.07565*.
11. Gysi, D.M. & Nowick, K., 2020. Construction, comparison and evolution of networks in life sciences and other disciplines. *Journal of the Royal Society Interface*, 17(166), 20190610.
12. Hartle, H., Klein, B., McCabe, S., Daniels, A., St-Onge, G., Murphy, C., & Hébert-Dufresne, L., 2020. Network comparison and the within-ensemble graph distance. *Proceedings of the Royal Society A*, 476: 20190744.
13. Hoel, E., Albantakis, L., & Tononi, G., 2013. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49), 19790-19795.
14. Klein, B., Hoel, E., Swain, A., Griebenow, R., & Levin, M., 2021. Evolution and emergence: Higher-order information structure in protein interactomes across the tree of life. *Integrative Biology*. 1-12. doi: 10.1093/intbio/zyab020.
15. Klein, B. & Hoel, E., 2020. The emergence of informative higher scales in complex networks. *Complexity*, 2020.
16. Newman, M.E., 2018. Network structure from rich but noisy data. *Nature Physics*, 14(6), 542-545.
17. Pascual, M. & Dunne, J.A. (Eds.), 2006. *Ecological networks: Linking structure to dynamics in food webs*. Oxford University Press.
18. Pinter-Wollman, N., Hobson, E.A., Smith, J.E., Edelman, A.J., Shizuka, D., De Silva, S., Waters, J.S., Prager, S.D., Sasaki, T., Wittemyer, G., Fewell, J., & McDonald, D.B., 2014.

- The dynamics of animal social networks: Analytical, conceptual, and theoretical advances. *Behavioral Ecology*, 25(2), 242-255.
19. Rosas, F.E., Mediano, P.A., Jensen, H.J., Seth, A.K., Barrett, A.B., Carhart-Harris, R.L., Bor, D., 2020. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, 16(12), e1008289.
 20. Shaw, J.O., Coco, E., Wootton, K., Daems, D., Gillreath-Brown, A., Swain, A., & Dunne, J.A. Disentangling ecological and taphonomic signals in ancient food webs. *Paleobiology*, 1-17.
 21. Puga-Gonzalez, I., Sueur, C. and Sosa, S., 2021. Null models for animal social network analysis and data collected via focal sampling: Pre-network or node network permutation?. *Methods in Ecology and Evolution*, 12(1), pp.22-32.
 22. Sporns, O., 2014. Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience*, 17(5), 652.
 23. Swain, A., Devereux, M., & Fagan, W.F., 2021. Deciphering trophic interactions in a mid-Cambrian assemblage. *iScience*, 102271.
 24. Swain, A., Williams, S.D., Di Felice, L.J., Hobson, E.A., 2021. Interactions, information and emergence: Exploring task allocation in ant colonies using network analysis. *bioRxiv*. <https://doi.org/10.1101/2021.03.29.437501>.
 25. Tononi, G., Sporns, O., & Edelman, G.M., 1999. Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, 96(6), 3257-3262.
 26. Tsimring, L.S., 2014. Noise in biology. *Reports on Progress in Physics*, 77(2), 026601.

27. Vázquez, D.P., Blüthgen, N., Cagnolo, L., & Chacoff, N.P. (2009). Uniting pattern and process in plant-animal mutualistic networks: A review. *Annals of Botany*, 103(9), 1445-1457.
28. Wang, Z., Andrews, M.A., Wu, Z.X., Wang, L., & Bauch, C.T., 2015. Coupled disease-behavior dynamics on complex networks: A review. *Physics of Life Reviews*, 15, 1-29.
29. van der Marel, A., Prasher, S., Carminito, C., O'Connell, C.L., Phillips, A., Kluever, B.M., Hobson, E.A., 2021. A framework to evaluate whether to pool or separate behaviors in a multilayer network. *Current Zoology*, 67(1), 101–111.